Employee Salary Prediction

Bavya Sri Vemulapalli
11733444
Department of Computer Science
University of North Texas

Jaswanth Sai Donepudi
11712789
Department of Computer Science
University of North Texas

Nikhil Kuchipudi
11709126
Department of Computer Science
University of North Texas

Nithish Kumar Boggula
11559328
Department of Computer Science
University of North Texas

Sai Sathwika Garimella
11595361
Department of Computer Science
University of North Texas

***Abstract*—This paper describes employee salary prediction, which plays a prominent role for any employee in the competitive world. Nowadays many factors affect one's salary, and to overcome these employees should prove themselves more productive. But there are some cases, where the employee is productive but still unable to attain their part of salary. So, these linear regression models help employees know their predicted salary and make them satisfy. The amount is predicted based on various features like age, experience, position, etc. So, we have taken five different linear regression models to enhance employee satisfaction by predicting employee salary.**

***Index Terms*—component, formatting, style, styling, insert**

I. Goals and Objectives

A. *Motivation*

The main motivation of this project is to help people understand the average salaries they get for a position in the market based on their human factors such as their age, their experience, the years they worked in the current role. These factors very much influence the salaries they could get if they change their company for the same position or if they want to change their role later in time. People do not know the average salaries they could expect for their position in the market. They are making mistake of working for a little wage because of the lack of knowledge. This is the main motivation of this project to make people aware and better understand the job market conditions and salaries.

B. *Significance*

The project is very significant among working class people and students who want to choose a career path. Students can make use of this project to see jobs in which fields are getting more packages thus they can take good courses and settle well in their lives. Working class people can use this project to observe the relations between different human factors and the salaries they could get. We used various machine learning models like linear regression, Gaussian mixture model, Random Forest regressor model, Gradient boosting regressor model, Bagging regressor model to make predictions of the salary.

C. *Objectives*

The primary objective of this project is to train different machine learning models on employee dataset. The project identifies renumeration offered by employers based on a holistic approach. The overall profile of employees is considered in this project to train the models. The main objective is to take input of employee details and give them a figure of expected salary based on our trained models.

D. *Features*

The dataset we took for this project has a lot of features which we will use initially to make different plots and give an idea of the dataset and then we take only the features which mostly impact the salary employee could get. The features we took for this project include Age, Total working years, Job role, Job level which are used to estimate monthly income of people. The figure below shows all the features in our dataset.

II. Related Work

A. *Employee Salary Analysis*

Employee salary analysis is very useful for employees as well as companies. Companies can use this project to analyse salaries given in market and employees can use this to negotiate with human resources.

B. *Machine Learning Models*

Machine learning can be used to predict salary a person could get based on the data we feed to the models. The project comes under supervised learning and regression problem in which we basically train a model by giving it labelled past data and then try to make predictions based on the trained data.

C. *Regression Models*

There are various types of regression models in which we take some models train them with a dataset from Kaggle and then see their results using evaluation metrics like mean squared error, mean absolute error and r-squared score. The models we used in our project are linear regression model, Gaussian mixture model, Random Forest regressor model, Gradient boosting regressor model and Bagging regressor model.

III. Dataset

We are going to work with the salary dataset which we have taken from Kaggle and it contains various fields like age, gender, education, job title, years of experience, salary. The dataset is Employee attrition Dataset. This dataset is taken from Kaggle and contains around 35 columns and 1470 rows and the type of the dataset is csv. The Employee Attrition dataset contains various attributes of data like age, attrition which shows if the employee will leave the company or not, department, education, gender, monthly salary, total working years etc. We use the required attributes to train our model and remove some unnecessary columns like over18, daily rate, overtime and other unimportant columns. Preprocessing steps include removing these unnecessary columns and encoding columns with string data into numerical format and see if there are any missing values or invalid data and change them if there are any. The dataset contains 26 integers, 6 string and 3 Boolean data type columns.
https://www.kaggle.com/datasets/patelprashant/employee-attrition

IV. Detail design of Features

There are a total of 35 columns. The detailed description of them is given below. Among these columns as we can see there are 9 other columns which are string columns labelled as objects.

They are Attrition, BusinessTravel, Departmment, EducationField, Gender, JobRole, MaritalStatus, Over18, OverTime. Remaining all other columns are interger columns represented as int64. There are no null items in the columns. There are no missing values in the dataset. All of the features are basically different properties of employees. The features include both personal and professional properties of employees.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64
 6   Education                 1470 non-null   int64
 7   EducationField            1470 non-null   object
 8   EmployeeCount             1470 non-null   int64
 9   EmployeeNumber            1470 non-null   int64
 10  EnvironmentSatisfaction   1470 non-null   int64
 11  Gender                    1470 non-null   object
 12  HourlyRate                1470 non-null   int64
 13  JobInvolvement            1470 non-null   int64
 14  JobLevel                  1470 non-null   int64
 15  JobRole                   1470 non-null   object
 16  JobSatisfaction           1470 non-null   int64
 17  MaritalStatus             1470 non-null   object
 18  MonthlyIncome             1470 non-null   int64
 19  MonthlyRate               1470 non-null   int64
 20  NumCompaniesWorked        1470 non-null   int64
 21  Over18                    1470 non-null   object
 22  OverTime                  1470 non-null   object
 23  PercentSalaryHike         1470 non-null   int64
 24  PerformanceRating         1470 non-null   int64
 25  RelationshipSatisfaction  1470 non-null   int64
 26  StandardHours             1470 non-null   int64
 27  StockOptionLevel          1470 non-null   int64
 28  TotalWorkingYears         1470 non-null   int64
 29  TrainingTimesLastYear     1470 non-null   int64
 30  WorkLifeBalance           1470 non-null   int64
 31  YearsAtCompany            1470 non-null   int64
 32  YearsInCurrentRole        1470 non-null   int64
 33  YearsSinceLastPromotion   1470 non-null   int64
 34  YearsWithCurrManager      1470 non-null   int64
dtypes: int64(26), object(9)
memory usage: 402.1+ KB
```

Fig. 1. Properties of features in Dataset

V. Analysis

- Import the required libraries like pandas, numpy, linear regression, random forest regression, mean squared error, r2 score, seaborn and matplotlib. Now, we need to read the dataset $employee-dataset.csv$ into the pandas data frame. Further, display the first few rows of the data frame using the head method. It displays columns like age, attribution, business travel, daily rate, department, education etc.
- Data info method is used to find out the detailed notes of the dataframe like non null count, datatype and memory usage.Data of a particular column with the head method is used to display the initial column values. Here we have displayed monthly income with integer datatype.Data of the particular column with describe method is used to define the descriptive statistics of the data frame. It helps us to find the mean, standard deviation, median etc. for each column.Data of a particular column with a unique method extracts the unique value in that column and prints it.Now, we need to display the heat map with the specific columns. Calculate the correlation matrix for the selected columns. Select the size of the heat map that needs to be displayed. Title the heatmap and display the heat map using Seaborn.
- Now, we need to create the pair plot for the following data related to monthly income and select the features like age, years in current role, years at the company, total working years, and monthly income. So different bar plots will be displayed for the selected features. The range of different monthly income is set the different coloured plots.
- Now, we need to set the boxplot using seaborn so to do that we need to set the size of the box plot, Set the x axis to Department and Y axis to MonthlyIncome. Finally, the title is set to boxplot of monthly income by attrition. These Departments are sales, research and development, and human resource.
- Now, we need to set the count plot to visualize the distribution of job roles to different levels of Education

in your data frame. Set the size of the figure for the count plot as said to represent the count on the X axis and Levels of education like sales executive, research scientist, Laboratory technician, manufacturing director, manager, and human resources.

- Now, we need to specify the columns you want to include in the pairplot and create a subset of the dataframe containing selected columns like Monthlyincome, Yearatcompany, job role. In which job role has multiple subfields. This function creates dummy variables for the job role in the dataframe with specific columns to be one hot encoded.
- In the Linear regression model, Take all columns of the dataframe except monthly income on X and monthly income is taken on Y. Make the training and testing sets from X and Y. Create a linear regression model, Train the model and get coefficients values and intercept values.
- The error value is displayed 1517.8288473926373 and R squared score is 0.8945888203249073. Plot the values in the scatter plot by taking the years of experience on the x label and salary on the y label. Title it as the Actual vs predicted salaries.
- Find out the linear error by subtracting predicted values from actual values. Visualize the distribution of errors using a histogram with kernel density estimate. Then find out the Mean squared error i.e. 2303804.409977, Mean Absolute error i.e. 1168.18985226073667 and R squared i.e. 0.8945888203249073. On X axis Errors and on the Y axis Frequency are displayed.
- Gaussian mixture model is created with two components, train the model with total working years. Then predict the values and store into labels and then create the two linear regression models. Now reshape the data into 2D array, Train both models, and plot the regression lines on scatterplot of data.
- In Random Forest regressor model, create a column transformer for preprocessing and split the data into training and testing sets. Preprocess the data and perform hyperparameter tuning to get the best parameters. Make the best predictions and finally, print best parameters. Mean squared error is 1206334.114593 Finally, plot the residual plot.
- Gradient boosting regressor model is used to create an object. Train the model and predict the values. Evaluate the model based on mean squared error, mean absolute error and R squared score. Finally, Plot the predicted vs actual values with a regression line and calculate the error metrics for gradient boosting.
- In Bagging regressor model, create a decision tree regressor object and bagging model. Train the model and make predictions. Visualize the actual vs predicted values and calculate the metrics. So that we can print metrics and runtime. In it Mean squared error i.e. 1370070.547, Mean Absolute error i.e. 877.007 and R squared i.e. 0.9373.

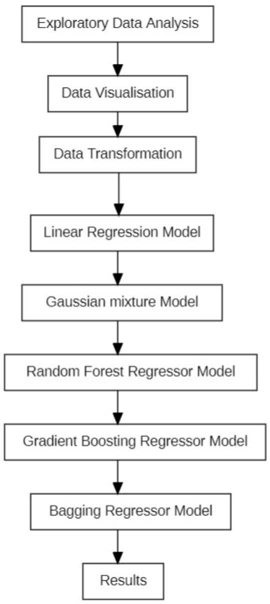

Fig. 2. Employee salary prediction Flowchart

VI. Implementation

A. Exploratory Data Analysis

In this phase dataset is read and studied which lets us know the type of data and different properties of our data. We get to know mean, std, min, max, median and other statistical properties of columns in our data using describe and info functions, and we can also slice the dataset and take our required columns and create new data frames from that.

B. Data Visualisation

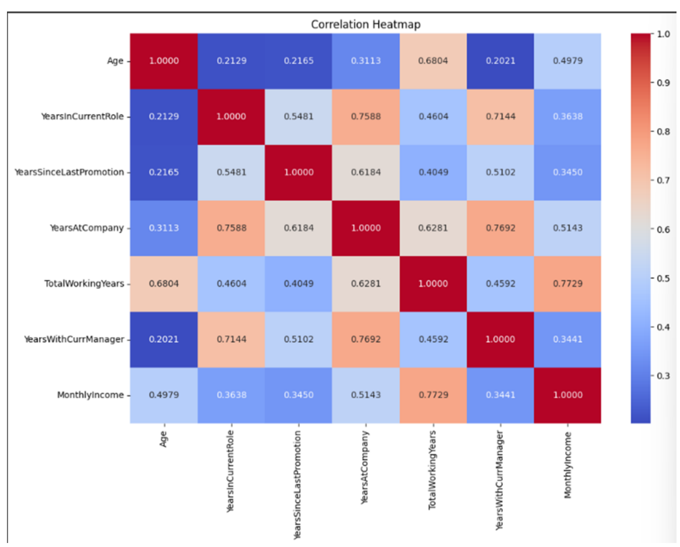

Fig. 3. Heat Map

In this phase we make different plots like Heatmap, Pairplot, Boxplot, Countplot, Scatterplot, Lineplots using matplotlib and

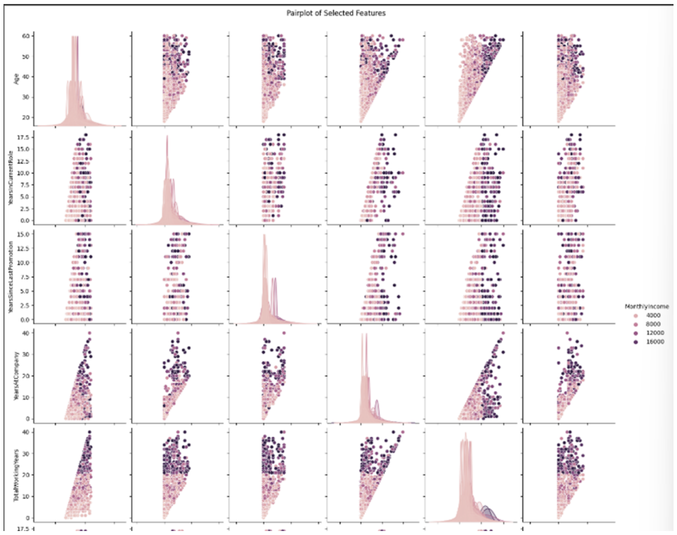

Fig. 4. Pairplot

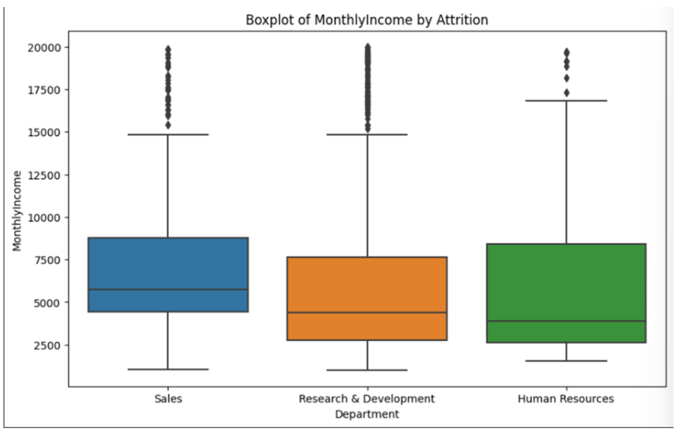

Fig. 5. Boxplot

seaborn. These help us understand our data visually in a pictorial form. Visualisation helps us a lot where plain data cannot. Plain data can be overwhelming and not so friendly to understand and draw conclusions. This is why we made some visualisations to visualise our data and show relations in it.

C. *Data Transformation*

We transform our data to required format. This includes removing unwanted columns which does not have any collinearity with salary variable. Transformation also includes encoding categorical values to numeric values using label encoder.

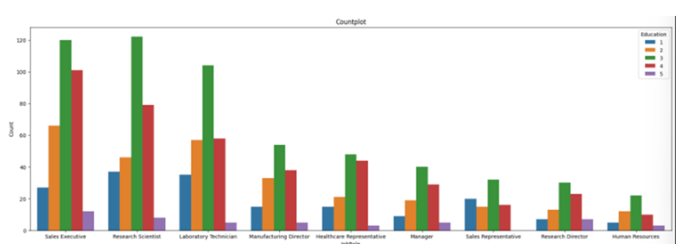

Fig. 6. Countplot

D. *Linear Regression model*

Linear regression is basically predicting the value of a dependent variable value based on single or multiple independent variables. The linear regression equation is given as

$Y \;=\; a1x1 \,+\, a2x2 \,+\, a3x3\ldots\ldots\ldots\ldots\ldots \,+\, anxn \,+\, b$

Here, Y is dependant variable and x1,x2. . . .n are independent variables and a1,a2. . . ..an are respective coefficients and b is the intercept value.

E. *Gaussian Mixture model*

Gaussian mixture model is a regression model which identifies underlying components as gaussian components in the data.

F. *Random Forest Regression model*

Random forest regression model is regression model which creates an ensemble of models learning. By analysing and averaging the predictions of multiple individual trees it builds the final response. By merging the predictions of decision trees, generalization is improved and overfitting will be reduced.

$g(x) = f0(x) + f1(x) + f2(x) + ...,$

where g denotes the final model which is sum base models.

G. *Gradient Boosting Regression model*

Similar to Random forest regression model, this also creates an ensemble of models learning. Here, an additive model is built in a repetitive process to add weak learners in order to reduce the loss function. This model primarily focuses on solving errors produced by previous models so that it becomes powerful but it weakens the system by overfitting.

$h_m(x_i) = y_i - F_m(x_i)$

H. *Bagging Regression model*

Similar to Random forest regression model, this also ensembles learning. By using different models on multiple data subsets, this method compresses overfitting. By averaging predictions, this method improves stability by reducing variance when merged with diverse models.

VII. Preliminary results

Linear Regression:

The coefficient values are -8.162, 50.592, -9.238, 3814.576 and the intercept value is -1574.524.

The root mean squared error value is 1517.8288.

The r-squared score value is 0.8945

The scatterplot of actual and predicted values and, the distribution of errors is given below.

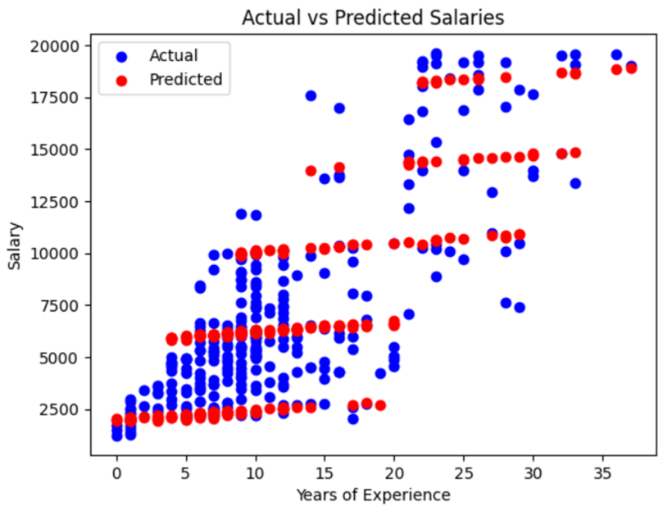

Fig. 7. Scatter Plot of Linear Regression

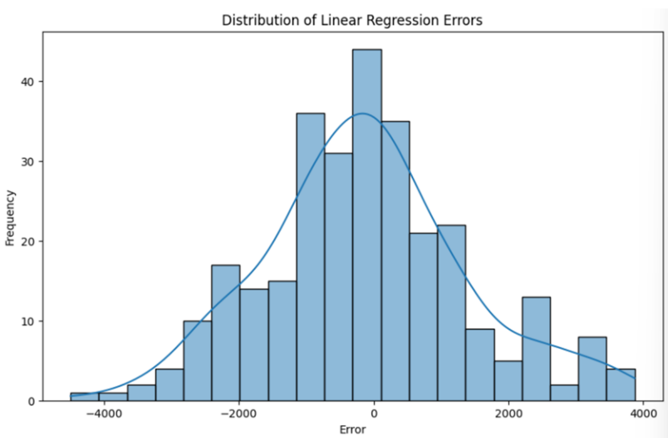

Fig. 8. Distribution of Errors

Gaussian Mixture Model:
The scatterplot and regression lines of the model is given below fig:9

Random Forest Regression Model:
The scatterplot and regression lines of the model is given below fig:10
Mean Squared Error: 1206334.114593432.
R-squared value is 0.9448

Gradient Boosting Regressor Model:
The scatterplot and regression lines of the model is given below fig:11
Mean squared error is 1300098.826
Mean absolute error is 838.0666
R-squared 0.940
The actual vs predicted salaries table is given below.

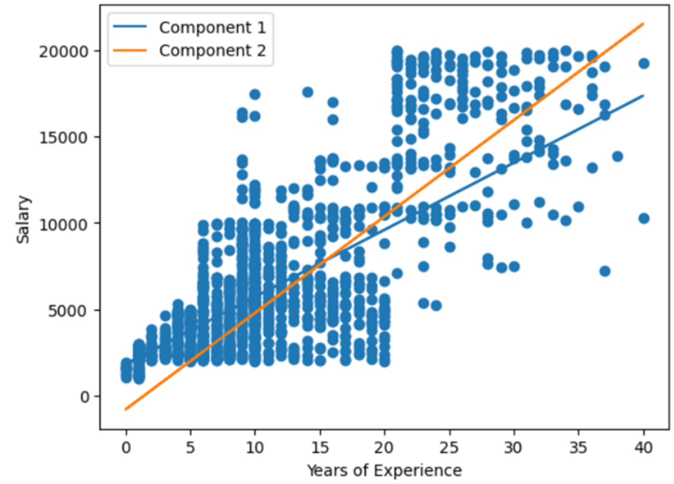

Fig. 9. Scatter plot of Gaussian Mixture Model

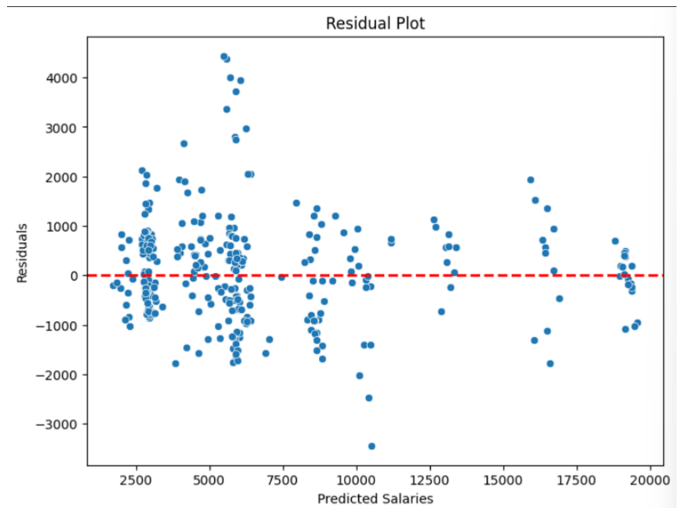

Fig. 10. Scatter plot of Random Forest Regression model

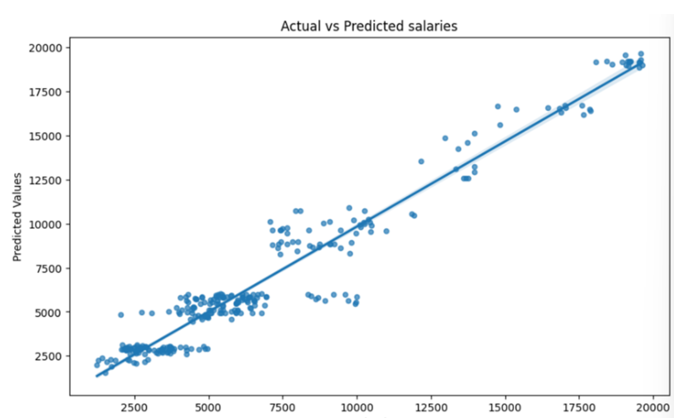

Fig. 11. Scatter plot of Gradient Boosting Regressor Model:

The distribution of errors is given below fig:12

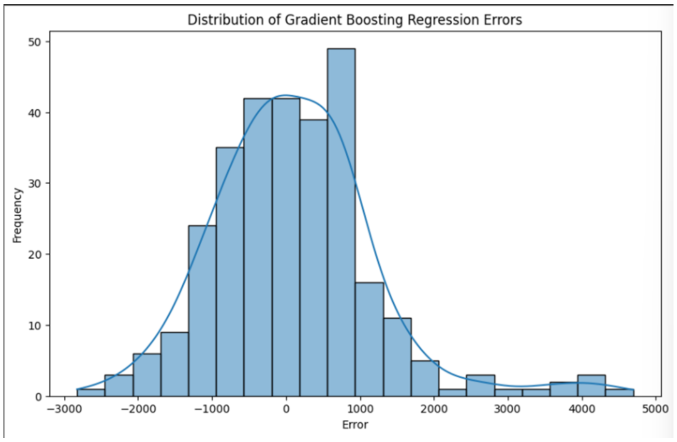

Fig. 12. Distribution of errors in Gradient Boosting Regressor Model:

Bagging Regression model:
The mean squared error value is 1370070.547
The mean absolute error value is 877.007
The r-squared score is given as 0.9373.
The actual vs predicted salaries regressor line is given below.

The distribution of errors is given below fig:13

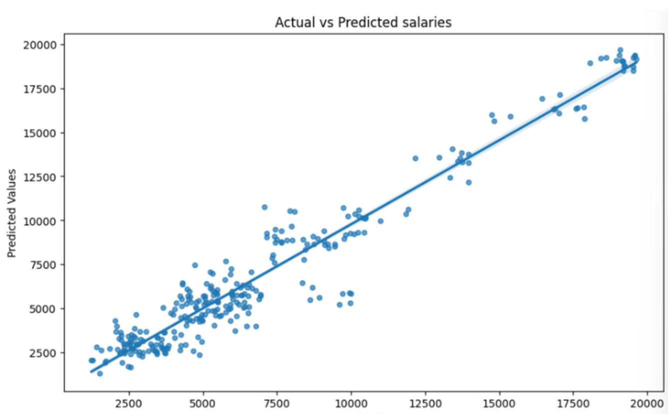

Fig. 13. Distribution of errors in Bagging Regression Model:

VIII. Project Management

Work completed:

- **Description:**
1. We have implemented till now 5 different regression and ensemble learning algorithms mentioned above. A good documentation is also done till now and the code and the documentation are uploaded to the github and the link is given in the documentation.
2. The error values of different evaluation metrics are also calculated and plots of error values and distribution of error terms are drawn using matplotlib and seaborn. We can say that almost 70 percent of our work is done till now.

- **Responsibility (Task, Person):**
1. Dataset collection, Base papers (IEEE papers) – Jaswanth, Bavya.
2. Data Pre-processing, Data visualisations - sathwika, Jaswanth, Bavya.
3. Machine learning models Implementation – Nikhil, Nithish.
4. Documentation – Sathwika, Nikhil, Jaswanth, Nithish.
5. Github Account – Nithish.

- **Contributions (members/percentage):**
Nikhil – 22
Nithish – 22
Jaswanth – 20
Bavya – 18
Sathwika - 18

Work completed:

- **Description:**
1. The only parts remaining in our project are implementing taking user input of variables and calculating their expected salary and the spss part.
2. The documentation also needs to be slightly tweaked to be more professional-looking and the code in the ipynb also needs to be updated with all evaluation metrics for all the models which are missing for some models in the current draft. All this work amounts close to 30 percent which when added would sum up to be 100 percent completion of our project.

- **Responsibility (Task, Person):**
1. Implementing code part of user salary prediction – Sathwika, Bavya, Jaswanth.
2. Spss part – Jaswanth, Nikhil
3. Documentation tweaking – Jaswanth, Bavya, Sathwika
4. Making IPYNB complete – Nithish, Nikhil

- **Contributions (members/percentage):**
Nikhil - 20
Nithish - 20
Jaswanth - 20
Bavya - 20
Sathwika - 20

IX. Project Github Link

https://github.com/Nithish-kumar-11/Group20_project

Acknowledgment

We would like to thank all the people who contributed to the success of the project. We would like to express our gratitude to professor Dr.Sayed Khushal Shah, Clinical Assistant Professor, Department of Computer science, University of North Texas for providing his continuous support throughout the project.

References

[1] S. S. Alduayj and K. Rajpoot, ”Predicting Employee Attrition using Machine Learning,” 2018 International Conference on Innovations in Information Technology (IIT), Al Ain, United Arab Emirates, 2018, pp. 93-98, doi: 10.1109/INNOVATIONS.2018.8605976.

[2] P. Viroonluecha and T. Kaewkiriya, ”Salary Predictor System for Thailand Labour Workforce using Deep Learning,” 2018 18th International Symposium on Communications and Information Technologies (ISCIT), Bangkok, Thailand, 2018, pp. 473-478, doi: 10.1109/ISCIT.2018.8587998.

[3] J. Vemulapati, A. Bayyana, S. H. Bathula, S. Tokala, K. Hajarathaiah and M. K. Enduri, ”Empirical Analysis of Income Prediction Using Deep Learning Techniques,” 2023 IEEE International Students’ Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 2023, pp. 1-6, doi: 10.1109/SCEECS57921.2023.10062992.