# 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

In this project we will predict the employee salary based on the following five different models like linear regression, Gaussian mixture, random forest regressor, Gradient boosting regressor and Bagging regressor models. Where we can predict the salaries of the employees. We can calculate it based on entering the employee age, total working years, job level and their job role. So that it will predict the monthly income of the employee. We have done data visualization before implementing the models. Related papers used only linear regression for salary prediction. We evaluated our models using different metrics and we finally calculated the salary of the employee based on the trained linear regression model.

## II. Related Work

In this case of predicting the salary of the employee, we need to make sure of giving the perfect trained model. Here, we have given five different models, which will predict the salary of the employee. Based on the data set we have taken there are high chances for predicting salary with high accuracy. Based on the literature review made on the similar papers, we found that this project is done through only linear regression. Here there are four more different algorithms that actually predicts the salary of the employee with best accuracies. We have taken monthly income as the dependent variable and age, job role, total working years, and job level are the rest of the features used to calculate the salary. Firstly, we found that the model is done with the help of small dataset i.e. 6 columns and 200 rows. So, we took a dataset with 35 columns and 1470 rows of data. Which will increase the efficiency of the model. Among those 35 columns we took required 5 columns

for executing the models. We need to perform exploratory data analysis. Which means we need to find out the relationships and hypothesis of the features. Now, we need to perform data visualisation techniques like heat map, pair plot, box plot and count plot. Data is transformed and linear regression model is done further. Now, we need to continue the rest of the four model i.e. gaussian mixture, random forest regressor, gradient boosting regressor and bagging models are performed. Now we need to predict the salary by using the linear regression model by taking manual inputs from the employees. Based on the data set the efficiency of the model increases.

## 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.

## 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.

```
<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

- 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.

### A. *SPSS Analysis*

**Descriptive Statistics:** Descriptive statistics are used to describe the main features of the dataset, this analysis tell us the way to organize large amount of data through mean, median and standard deviation. Mean age is 36.92 years and Standard Deviation is 9.135. Monthly income 1ranges from 1009 to 19999 and mean is 6502.93 with standard deviation of 4707.957. Job Level ranges from 1 to 5, a mean of 2.06 and a standard deviation of 1.107. Total Working Years with a mean of 11.28 years and standard deviation of 7.781.

**Descriptive Statistics**

| | N | Minimum | Maximum | Mean | Std. Deviation | Variance |
|---|---|---|---|---|---|---|
| Age | 1470 | 18 | 60 | 36.92 | 9.135 | 83.455 |
| MonthlyIncome | 1470 | 1009 | 19999 | 6502.93 | 4707.957 | 22164857.072 |
| JobLevel | 1470 | 1 | 5 | 2.06 | 1.107 | 1.225 |
| TotalWorkingYears | 1470 | 0 | 40 | 11.28 | 7.781 | 60.541 |
| Valid N (listwise) | 1470 | | | | | |

Fig. 2. Descriptive Statistics

**Correlations:** Correlation matrix projects the relationship between variables like age, job level, total working years with monthly income. They depict the interdependencies of factors. Correlation is the statistical measure describes the change in two variables. Correlation coefficient is a numerical value that range from -1 to 1. 1 represents positive correlation, 0 represents no systematic relation and -1 represents negative correlation. Correlation is significant at the 0.01 level.

**Correlations**

| | | Age | MonthlyIncome | JobLevel | TotalWorkingYears |
|---|---|---|---|---|---|
| Age | Pearson Correlation | 1 | .498** | .510** | .680** |
| | Sig. (2-tailed) | | <.001 | <.001 | <.001 |
| | N | 1470 | 1470 | 1470 | 1470 |
| MonthlyIncome | Pearson Correlation | .498** | 1 | .950** | .773** |
| | Sig. (2-tailed) | <.001 | | <.001 | <.001 |
| | N | 1470 | 1470 | 1470 | 1470 |
| JobLevel | Pearson Correlation | .510** | .950** | 1 | .782** |
| | Sig. (2-tailed) | <.001 | <.001 | | <.001 |
| | N | 1470 | 1470 | 1470 | 1470 |
| TotalWorkingYears | Pearson Correlation | .680** | .773** | .782** | 1 |
| | Sig. (2-tailed) | <.001 | <.001 | <.001 | |
| | N | 1470 | 1470 | 1470 | 1470 |

**. Correlation is significant at the 0.01 level (2-tailed).

Fig. 3. Correlations

**Regression Analysis:** We can see that Total working years and age have higher positive effect and job level has somewhat

less positive effect. Here, Monthly Income is the dependent variable and predictors are total working years, age, job level.

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .952[a] | .905 | .905 | 1449.227 |

a. Predictors: (Constant), TotalWorkingYears, Age, JobLevel

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 29481194269 | 3 | 9827064756.3 | 4678.976 | <.001[b] |
| | Residual | 3078980769.3 | 1466 | 2100259.733 | | |
| | Total | 32560175038 | 1469 | | | |

a. Dependent Variable: MonthlyIncome

b. Predictors: (Constant), TotalWorkingYears, Age, JobLevel

**Coefficients[a]**

| Model | | Unstandardized Coefficients B | Unstandardized Coefficients Std. Error | Standardized Coefficients Beta | t | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Constant) | -1621.307 | 178.919 | | -9.062 | <.001 |
| | Age | -7.581 | 5.655 | -.015 | -1.341 | .180 |
| | JobLevel | 3784.737 | 54.895 | .890 | 68.945 | <.001 |
| | TotalWorkingYears | 52.543 | 9.169 | .087 | 5.731 | <.001 |

a. Dependent Variable: MonthlyIncome

Fig. 4. Regression Analysis

**ANOVA:** We can see from the anova table that the model is a good fit for monthly income. The F-statistic value 4678.976 also supports the effectiveness of the model. The significance value is also less than 0.001

**Bayesian ANOVA Table:** The table gives the Monthly income for different factors given above. Factor analysis between age and monthly income is given as r = 0.865. We can see from the table that the age and monthly income captured a significant proportion of the total variance showing their relevance in our dataset.

**ANOVA**

| MonthlyIncome | Sum of Squares | df | Mean Square | F | Sig. | Bayes Factor[a] |
|---|---|---|---|---|---|---|
| Between Groups | 26570984803 | 8 | 3321373100.4 | 810.214 | <.001 | . |
| Within Groups | 5989190234.7 | 1461 | 4099377.300 | | | |
| Total | 32560175038 | 1469 | | | | |

a. Bayes factor: JZS

Fig. 5. Bayesian ANOVA

**Frequencies and Visualizations:** We have drawn the frequency tables, and other visualisations which provides additional info about age, Total working years, job level and job role factors which are used to train our model. Bar chart is plotted for age, Total working years, job role and job level with respect to frequency and their variables.

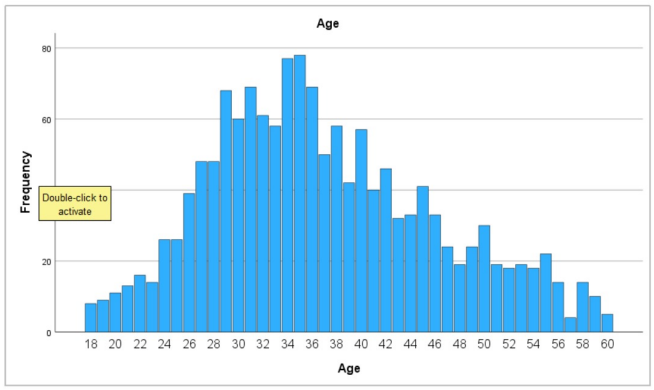


Fig. 6. Frequency bar chart for Age

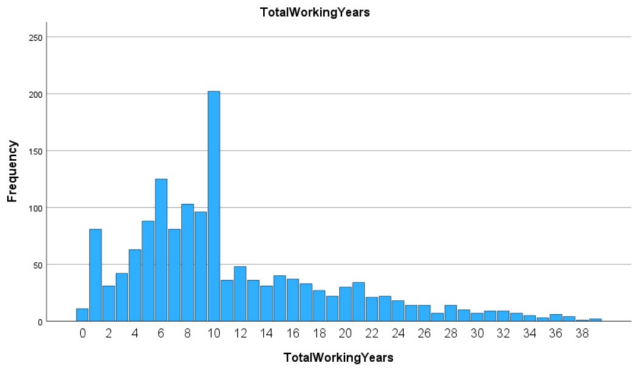


Fig. 7. Frequency bar chart for TotalWorkingYears

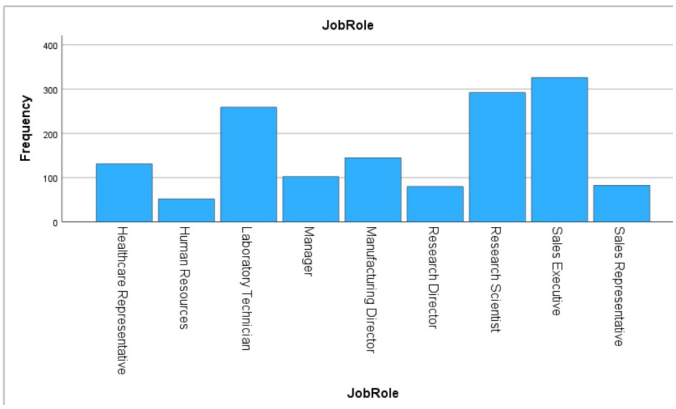


Fig. 8. Frequency bar chart for JobRole

## VI. Implementation

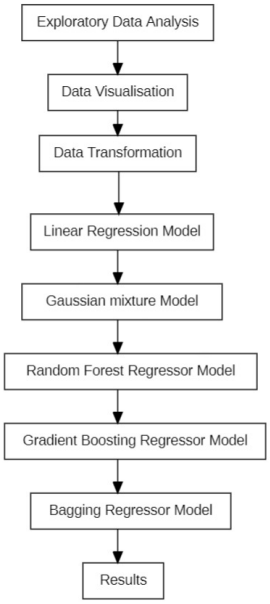


Fig. 9. Employee salary prediction Flowchart

### 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*

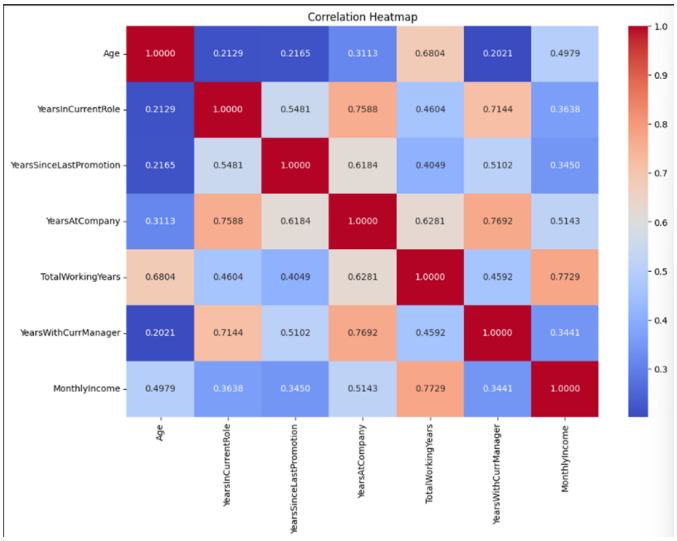


Fig. 10. Heat Map

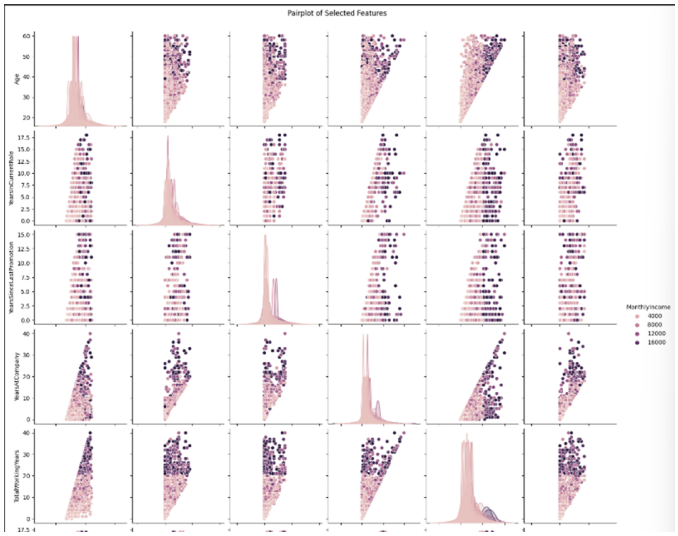


Fig. 11. Pairplot

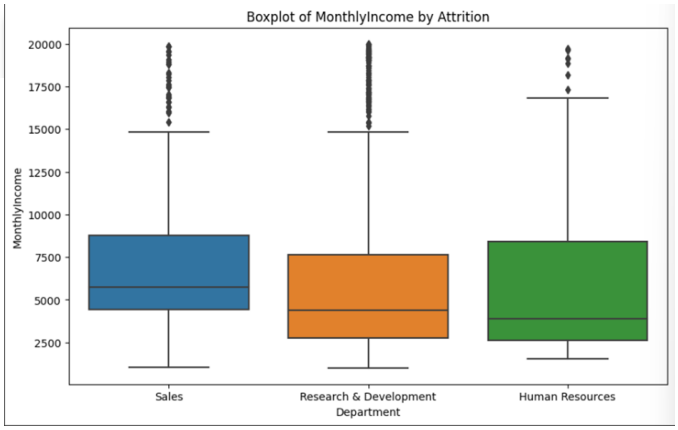


Fig. 12. Boxplot

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

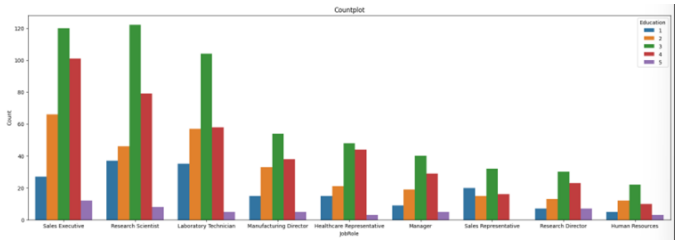


Fig. 13. Countplot

### 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.

### 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
$Y = a1x1 + a2x2 + a3x3..... + anxn + b$
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.

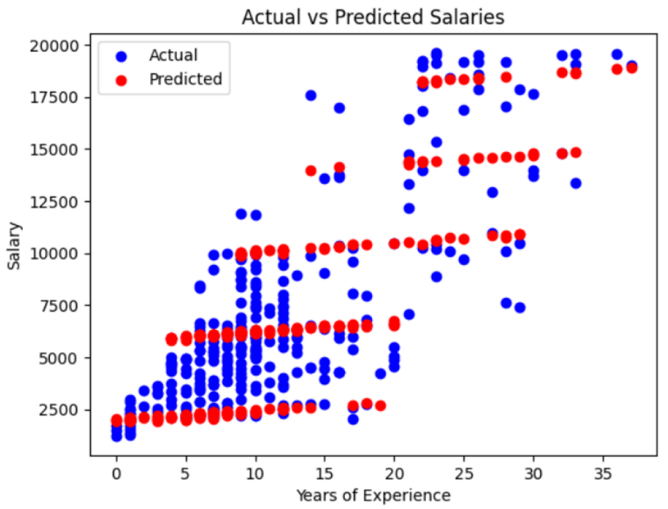


Fig. 14. Scatter Plot of Linear Regression

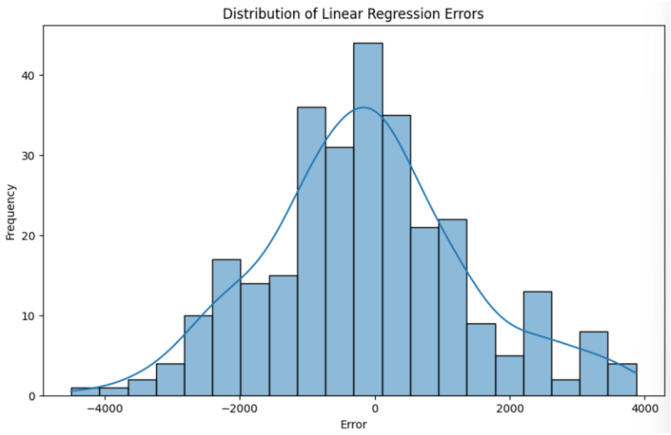


Fig. 15. Distribution of Errors

**Gaussian Mixture Model:**
The scatterplot and regression lines of the model is given below fig:16

**Random Forest Regression Model:**
The scatterplot and regression lines of the model is given below fig:17
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:18
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.

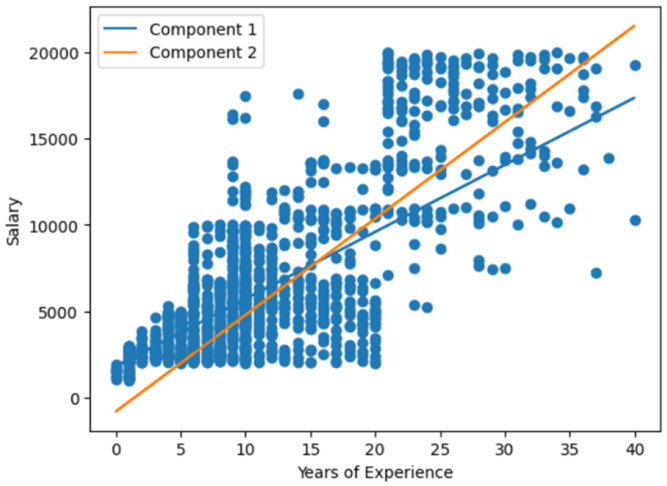


Fig. 16. Scatter plot of Gaussian Mixture Model

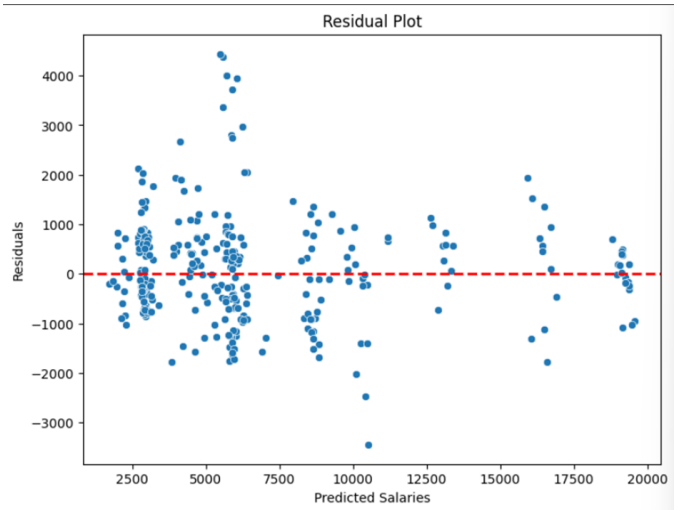


Fig. 17. Scatter plot of Random Forest Regression model

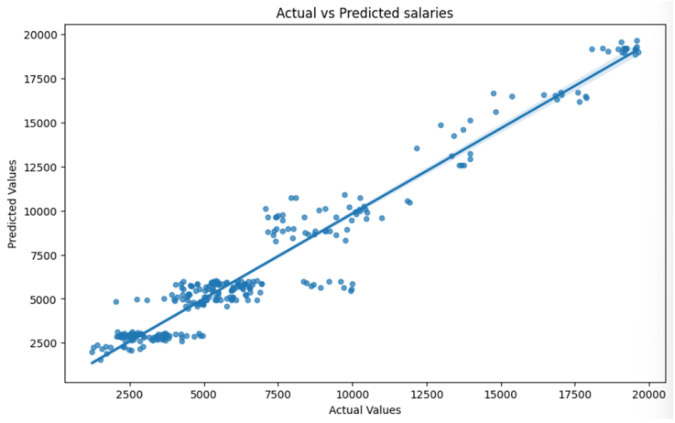


Fig. 18. Scatter plot of Gradient Boosting Regressor Model:

The distribution of errors is given below fig:19

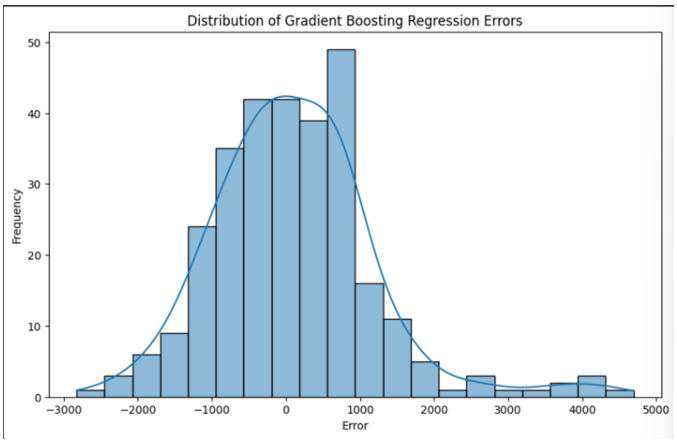


Fig. 19. Distribution of errors in Gradient Boosting Regressor Model:

**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:20

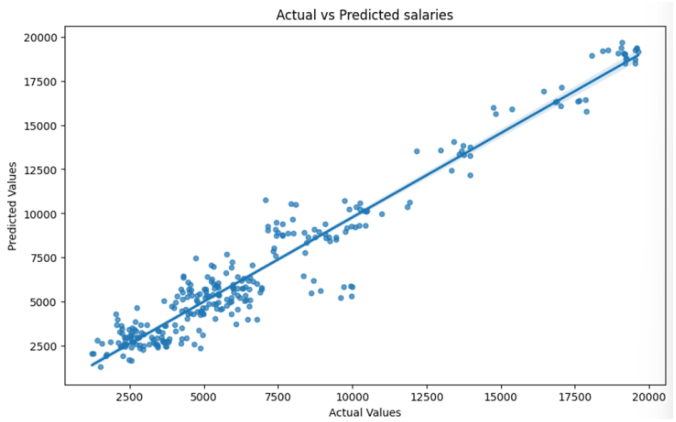


Fig. 20. Distribution of errors in Bagging Regression Model:

## VIII. Project Management

### A. *Implementation Status Report*

**Employee salary prediction using linear regression model:** Based on the coefficients and intercept values, we got from linear regression model. We formulate an equation to calculate the monthly salary of an employee by taking input values from the users like enter employee age, enter employee working years and the job level in the scale of 1 to 5. So that we can enter employee job role like healthcare representative as 0, human resources as 1, laboratory technician as 2, manager as 3, manufacturing director as 4, research director as 5, research scientist as 6, sales executive as 7, sales representative as 8. Here we need to enter the respective number in the space of enter employee job role. So that it will display the predicted monthly income of employee.

```
Enter Employee Age:
50
Enter Employee Total Working years:
30
Enter Employee Job Level(1-5):
4
Healthcare Representative: 0
Human Resources: 1
Laboratory Technician: 2
Manager: 3
Manufacturing Director: 4
Research Director: 5
Research Scientist: 6
Sales Executive: 7
Sales Representative: 8

Enter Employee Job Role:
6
Predicted monthly income of Employee is:  13388.00639 USD
```

Fig. 21. Employee salary prediction using linear regression model

**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 all of our work is done.
  3. We implemented taking user input of variables and calculating their expected salary and the spss analysis is performed.

- **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. Implementing code part of user salary prediction – Bavya, Jaswanth.
  5. Spss part – Jaswanth, Nikhil
  6. Documentation – Jaswanth, Bavya, Sathwika
  7. Github – Nithish, Nikhil

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

## 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.