SALARY PREDICTION USING LINEAR REGRESSION

Subhendu Mandal(23MA60R11), Susobhan Pratihar(23MA60R12), Md Bunty Ansari(23MA60R30), Rajdeep Thaosen (23MA60R35)

MTech, Department of Mathematics, IIT Kharagpur

ABSTRACT:

Machine Learning's (ML) ability to derive accurate outcomes without explicit programming has propelled the development of predictive engines. This project focuses on constructing a robust salary prediction model utilising key features essential for estimating employee salaries. Using Linear Regression as the chosen algorithm, this model aims to forecast salaries for graduates and visualise the findings through user-friendly graphs. The objective is to offer insights into salary growth within specific fields based on qualifications, facilitating informed decision-making for employers and contributing to a more data-driven approach in human resource management.

Key words: Salary Prediction, Regression, R-square.

1. INTRODUCTION :

A prediction is an assumption about a future event. A prediction is sometimes, though not always, based upon knowledge or experience. Future events are not necessarily certain, thus confirming exact data about the future is in many cases impossible, a prediction may be useful to help in preparing plans about probable developments. In this paper the salary of an employee of an organisation is to be predicted on the basis of previous salary growth rate. Here the history of salary has been observed and then on the basis of that salary of a person after a certain period of time it can be calculated automatically. In this paper the main aim is predicting salary and making a suitable user-friendly graph. From this prediction the salary of an employee can be observed according to a particular field according to their qualifications. It helps to see the growth of any field. It can produce a person's salary by clustering and predicting the salary through the graph. Using linear regression and polynomial regression it makes a graph. This graph helps to predict the salary for any position.

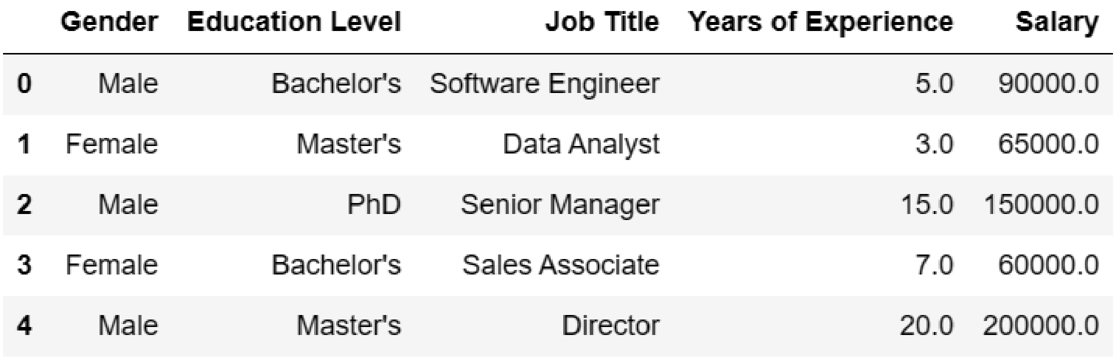
2. METHODOLOGY :

Data Collection: Gather a comprehensive dataset containing relevant features such as education level, years of experience, and job title , alongside corresponding salary information.
Data Preprocessing: Cleaned and prepared the data by handling missing values, encoding categorical variables, and scaling numerical features to ensure uniformity across the dataset.
Feature Selection: Identified key features impacting salary through exploratory data analysis and selected a subset of influential variables for model training.
Train-Test Split: Divided the dataset into training and testing sets to assess the models' generalisation performance accurately.
Linear Regression: Implemented a linear regression model to establish a baseline for salary prediction, considering the linear relationship between features and salary.

Model Training and Evaluation: Trained each model on the training set, fine-tuning hyperparameters, and evaluated their performance using metrics like Mean Squared Error (MSE) or R-squared on the test set.

3. DATA DESCRIPTION :

The dataset contains the record of 6704 employees, each of which has 5 features. i.e, there are 6704 entries with 5 features. Here the first five rows are given below.



	Gender	Education Level	Job Title	Years of Experience	Salary
0	Male	Bachelor's	Software Engineer	5.0	90000.0
1	Female	Master's	Data Analyst	3.0	65000.0
2	Male	PhD	Senior Manager	15.0	150000.0
3	Female	Bachelor's	Sales Associate	7.0	60000.0
4	Male	Master's	Director	20.0	200000.0

Fig-1: First five rows of the dataset.

Distribution of Categorical Variables:

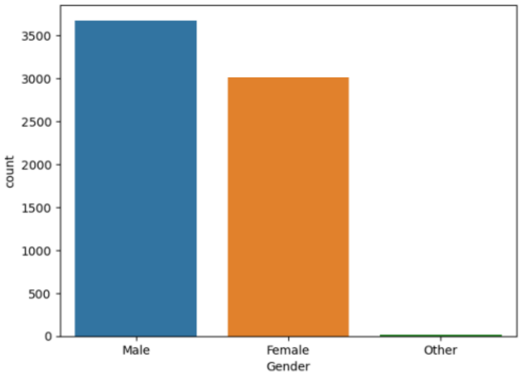


Fig-2

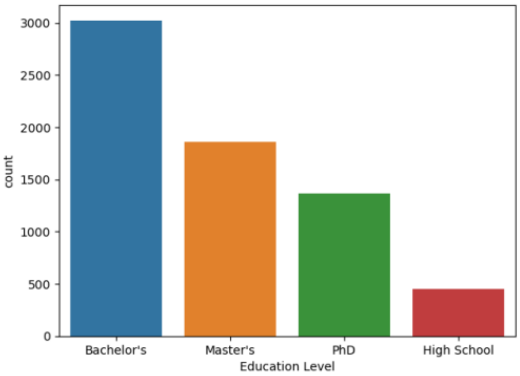


Fig-3

Fig-2 reveals that a significant portion of the employees are males, while Fig-3 indicates that the majority of employees have completed a bachelor's degree.

Distribution of Continuous variables:

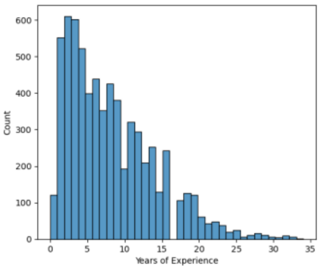


Fig-5

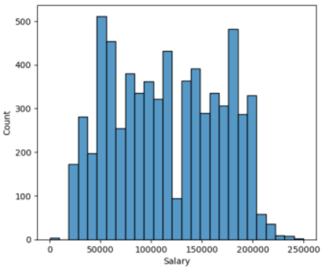
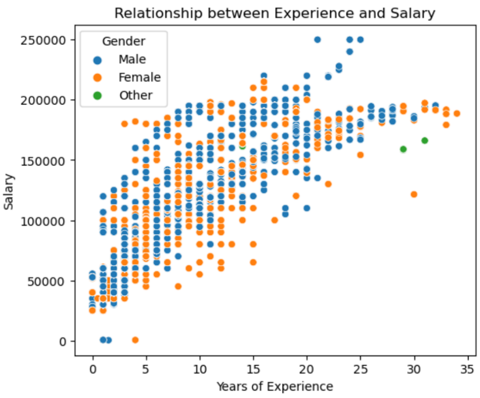


Fig-6

Fig-5 illustrates employees' experience levels with the majority having 1 to 10 years of experience, While Fig-6 demonstrates the salary distribution with most employees earning salaries between 50,000 to 2,00,000

Relationship with Target Variable:



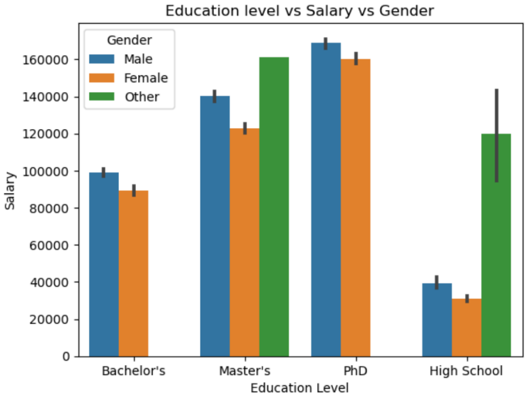


Fig-7 Fig-8

Fig-7 shows the relationship between experience and salary of employees. It illustrates that as experience increases salary also increases. Gender distributions are also the same.

And, Fig-8 shows education level and salary among the genders. In all education level categories male get a higher salary than females. In Master's and High School categories other gender get a higher salary than males and females.

Correlation Matrix:

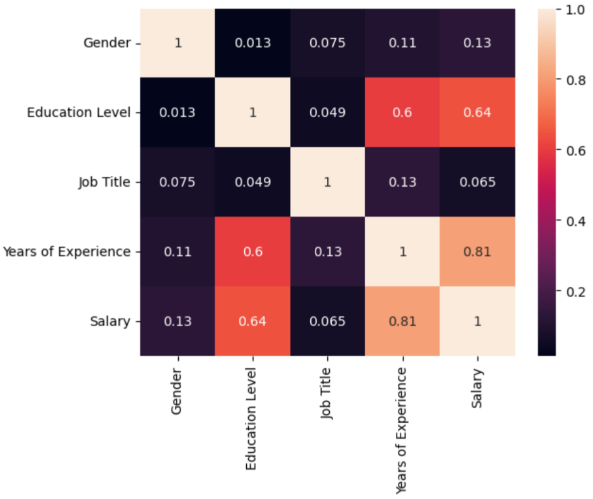


Fig-9

4.REGRESSION ANALYSIS:

The regression model is developed for predicting salary, considering factors contributing in salary increasing as independent variables and Salary as dependent variable using a regression equation.

A.Introduction:

The form for linear regression models developed will be in following form:

$y=mx+c$

Where,

y =Salary to be predicted.
x= factors contributing to predict the salary.
m & c *are* slope and intercept respectively.
For multiple regressions if there are n predictor variables, then the regression equation model is,
$y=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + e$

The $x_1, x_2, \ldots x_n$ represent the n predictor variables. Those parameters are the same as before, β_0 is the constant, β_1 is the coefficient on the first predictor variable,β_2 is the coefficient on the second predictor variable, and so on. e is the error term.

B. Steps for Regression Analysis:

1) There are a total 5 parameters that have been identified which are further subdivided into two categories.
   - Y(dependent variable)
     1.Salary
   - X(independent variables)
     1.Gender
     2.Education Level
     3.Job title
     4.Years of Experience
2) For a simple linear one independent variable(x), Years of Experience is taken to predict the Salary.
3) For multiple linear regression the following parameters are chosen on the basis of correlation and apply regression.
   - Gender
   - Education Level
   - Years of Experience

C. Application of Regression Analysis:

1.Simple Linear Regression:
y= Salary , x_1= Years of Experience

2. Multiple Linear Regression:
y=Salary, x_1=Gender , x_2=Education Level , x_3=Years of Experience.

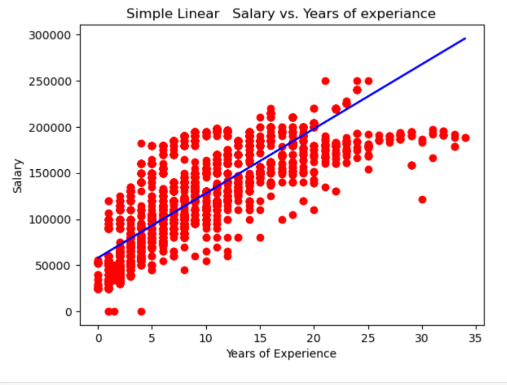


Fig-10

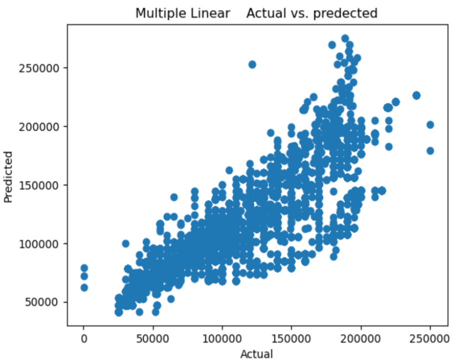


Fig-11

Fig-10 shows the best fitted curve for simple linear regression, while Fig-11 shows the scatterplot between actual salary and predicted salary in multiple linear regression.

Table 1 shows the results of application of regression

Table 1 : Summary Output

Regression	Statistics (Simple Linear Regression)	Statistic (Multiple Linear Regression)
R Square	.654	.695
Adjusted R Square	.654	.695
Observations	6698	6698

	Coefficients (Simple Linear Regression)	Coefficients (Multiple Linear Regression)
Intercept	58061.19643	41482.23698
x_1	6992.35166	6524.02064
x_2	-	15553.80789
x_3	-	5508.38536

From the regression analysis output we obtain coefficients for parameters and we obtained equation for simple linear regression is:

$y=58061.19643 + 6992.35166x$

Similarly, we obtained the equation for multiple linear regression is:

$y=41482.23698 + 6524.02064x_1 + 15553.80789x_2 +5508.38536x_3$

D. Prediction :

Using the equation from regression analysis of salary prediction is made for different given actual values. Error in respective model is identified and their validity is checked in Table 2 .

Table 2: Prediction for salary

Actual value of Salary	Predicted Value of Salary (Simple Linear)	Predicted Value of Salary (Multiple Linear)	Accuracy (%) (Simple Linear)	Accuracy(%) (Multiple Linear)
90000	93022.9	91102	96.64	98.78
140000	155954	156231	88.60	88.40
75000	93022.9	91102	75.97	78.53
62000	79038	64531	72.52	95.92

5. CONCLUSION:

This project extensively utilised machine learning techniques, specifically Linear Regression models, to forecast employee salaries based on diverse factors like experience, education level, gender, and job titles. Through a thorough analysis, it was evident that Multiple Linear Regression outperformed Simple Linear Regression, showcasing higher accuracy in predicting salaries due to its consideration of multiple predictors.

Insights extracted from exploratory data analysis shed light on significant relationships between experience, education, and gender with salary variations. Visual representations effectively illustrated these correlations, emphasising the impact of these factors on salary levels.

In conclusion, the project underscores the efficacy of employing Multiple Linear Regression for salary prediction, providing employers with a more comprehensive and accurate tool for decision-making in compensation planning. Future advancements may involve exploring advanced algorithms and refining feature engineering to further enhance the predictive capabilities.

6. REFERENCES:

[1] Lee, Y. and Sabharwal, M.: Education–Job Match, Salary and Job Satisfaction across the Public, Non- Profit and For-Profit Sectors: Survey of recent college graduates. Public Management Review 18 (1), 40-64 (2014).

[2] Jerrim, J.: Do College Students Make Better Predictions of Their Future Income Than Young Adults in the Labor Force? Education Economics 23(2), 162-179 (2013).

[3] Karla, R. H. and Hamlen, W. A.: Faculty Salary as a Predictor of Student Outgoing Salaries From MBA Programs. Journal of Education for Business 91(1), 38-44 (2015).

[4] Romero, C. and Ventura, S.: Educational Data mining: A Review of the State of the Art. IEEE Transactions on Systems Man and Cybernetics 40(6), 601-618 (2010).

[5] Singh, R.: A Regression Study of Salary Determinants in Indian Job Markets for Entry Level Engineering Graduates. Masters Dissertation. Dublin Institute of Technology (2016).

Appendix:

Method-1: Simple linear regression

```
# import needed libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error,mean_absolute_error
from sklearn.metrics import r2_score
import statsmodels.api as sm

# Load the data from the CSV file
salary_df = pd.read_csv('Salary_Data.csv')
#salary_df=salary_df.head(1000)
columns_to_drop = ['Gender', 'Education Level','Job Title']
# Drop the specified columns
salary_df = salary_df.drop(columns=columns_to_drop)

# detecting the null values
salary_df.isnull().sum()
from sklearn.impute import SimpleImputer
# drop the null values
salary_df.dropna(inplace = True)


X = salary_df['Years of Experience']
y = salary_df['Salary']


#dividing train and test.
X_train,X_test,y_train,y_test = train_test_split(X,y,train_size=0.2,random_state=42
X_train = np.array(X_train)[:, np.newaxis]
X_test = np.array(X_test)[:, np.newaxis]


lr = LinearRegression()
lr.fit(X_train,y_train)


# Predicting the Salary for the Test values
y_pred = lr.predict(X_test)

# Plotting the actual and predicted values
plt.scatter(X_test,y_test,color='r',linestyle='-')
plt.plot(X_test,y_pred,color='b',linestyle='-')
plt.xlabel('Years of Experience')
plt.ylabel('Salary')
plt.title('Salary vs. Years of experience')
plt.show()

lr.score(X_test,y_test)

y_pred_lr = lr.predict(X_test)


#print("Mean Squared Error :",mean_squared_error(y_test,y_pred_lr))
print("Mean Absolute Error :",mean_absolute_error(y_test,y_pred_lr))
print("Root Mean Squared Error :",mean_squared_error(y_test,y_pred_lr,squared=False
print("R-squared:", r2_score(y_test, y_pred_lr))

# Print the coefficients and intercept
slope = lr.coef_[0]
intercept = lr.intercept_
print(f"Slope (Coefficient): {slope}")
print(f"Intercept: {intercept}")

# Replace with actual values
new_data = pd.DataFrame({'Years of Experience': [5,3,15,14,20]})

predictions = lr.predict(new_data)
# Display the predictions
print("Predictions:")
print(predictions)


#Summary
model = sm.OLS(y, X).fit()
#view model summary
print(model.summary())


# Make predictions on the test data
y_pred = model.predict(X_test)
# Plot actual vs predicted values
plt.scatter(y_test, y_pred)
plt.xlabel("Actual Values")
plt.ylabel("Predicted Values")
plt.title("Actual vs Predicted Values in Linear Regression")
plt.show()
```

Method-2: Multiple linear regression

```
# import needed libraries
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error,mean_absolute_error
from sklearn.metrics import r2_score
from sklearn.feature_extraction.text import TfidfVectorizer
import statsmodels.api as sm

# read a csv file
salary_df = pd.read_csv('Salary_Data.csv')
salary_df.head()



salary_df.isnull().sum()
from sklearn.impute import SimpleImputer

salary_df.dropna(inplace = True)

salary_df

# # creating the variable for reducing the number of job titles
job_title_stats = salary_df['Job Title'].value_counts()
job_title_stats_less_than_50 = job_title_stats[job_title_stats<=50]
job_title_stats_less_than_50.count()

salary_df['Job Title'] = salary_df['Job Title'].apply(lambda x: 'Others' if x in
job_title_stats_less_than_50 else x)
salary_df['Job Title'].nunique()
salary_df



# checking unique values in education level
salary_df['Education Level'].unique()

salary_df['Education Level']=salary_df['Education Level'].replace(["Bachelor's
Degree","Master's Degree","phD"],["Bachelor's", "Master's", 'PhD'])

…
salary_df['Education Level'].unique()
```

```
salary_df.Gender.value_counts()

#Bar plot
fig, ax = plt.subplots(1,2,figsize=(15,5))
sns.countplot(x='Gender',data=salary_df,ax = ax[0])
sns.countplot(x='Education Level',data = salary_df,ax=ax[1])

# detecting the outliers in salary column using IQR method
Q1 = salary_df.Salary.quantile(0.25)
Q3 = salary_df.Salary.quantile(0.75)
IQR = Q3-Q1
lower = Q1-1.5*IQR
upper = Q3+1.5*IQR

salary_df

salary_df['Gender']=salary_df['Gender'].astype('category')
salary_df['Gender']=salary_df['Gender'].cat.codes
education_mapping = {"High School":0,"Bachelor's":1,"Master's":2,"PhD":3}
salary_df['Education Level'] = salary_df['Education Level'].map(education_mapping)
salary_df['Education Level']=salary_df['Education Level'].astype('category')
salary_df['Education Level']=salary_df['Education Level'].cat.codes
salary_df['Job Title']=salary_df['Job Title'].astype('category')
salary_df['Job Title']=salary_df['Job Title'].cat.codes
salary_df

salary_df[salary_df.Salary>upper]

salary_df[salary_df.Salary<lower]

# Correlation plot
sns.heatmap(salary_df.corr(),annot = True)

X = salary_df.drop(columns='Salary')
X = X.drop(columns='Job Title')
#X = sm.add_constant(X)
X.head()

X_train,X_test,y_train,y_test =
train_test_split(X,y,train_size=0.2,random_state=100)

lr = LinearRegression()
lr.fit(X_train,y_train)

c=lr.intercept_
```

```
print(c)

m=lr.coef_
print(m)

y_pred_test=lr.predict(X_test)
y_pred_test


plt.scatter(y_test,y_pred_test)
plt.xlabel('Actual')
plt.ylabel('Predicted')

#print("Mean Absolute Error :",mean_absolute_error(y_test,y_pred_test))
#print("Root Mean Squared Error :",mean_squared_error(y_test,y_pred_test,squared=Fa
print("R-squared:", r2_score(y_test, y_pred_test))

#fit linear regression model
model = sm.OLS(y, X).fit()
#view model summary
print(model.summary())
```