CTC Prediction Based On Various Factors Using Machine Learning

1 - Overview

I developed a CTC (Cost to Company) prediction model using three datasets: employee, college, and city data. The employee dataset included information about the employees such as their roles, salaries, colleges, cities, and other relevant details.

2 - Data Preparation

2.1 - College Dataset : Contained three columns representing tiers (1, 2, 3) with the names of colleges.

2.2 - City Dataset : Included two columns for metro and non-metro cities.

3 - Data Transformation

3.1 - Categorical to Numerical Conversion :

- Colleges were categorized as 1, 2, and 3 based on their tiers.

- Cities were encoded as 0 (non-metro) and 1 (metro).

3.2 - Data Merging :

The transformed college and city data were merged with the employee dataset.

4 - Data Preprocessing

4.1 - Outlier Removal : Identified and removed outliers to ensure data quality.

4.2 - Data Scaling : Normalized the data for consistent model training.

5 - Model Training

I trained three regression models:

1. Linear Regression

2. Decision Tree Regressor

3. Random Forest Regressor

6 - Model Evaluation

The models were evaluated based on their R^2 values.

Random Forest Regressor outperformed the others with the highest R^2 value.

7 - Further Improvements

Discussed strategies to enhance the model's performance and improve prediction accuracy.