Predicting Employee CTC Using Machine Learning

A Regression-Based Approach

- Meghana L

Problem Statement

To predict the Cost to Company (CTC) for employees based on factors like college tier, city type, role, experience, and previous CTC.

- **Challenges:**
  - Handling categorical variables.
  - Identifying non-linear relationships in the data.
  - Balancing model complexity and accuracy.

Dataset Overview

- **Data Sources:**
  - Employee data: [Number of rows]
  - College tiers: [Number of tiers]
  - City types: [Metro/Non-Metro categories]
- **Key Features:**
  - Categorical: Role, College Tier, City Type
  - Numerical: Experience, Graduation Marks, Previous CTC
- **Target Variable:** Cost to Company (CTC)

Data Preprocessing

1.Handling Missing Values:

- Checked and imputed or dropped null values.

2. Feature Transformation:

- Mapped colleges to tiers (1, 2, 3).
- Classified cities as metro (1) and non-metro (0).
- Converted roles to binary (Manager = 1, Executive = 0).

3.Feature Scaling:

- Applied standard scaling to numerical features to normalize them.

Exploratory Data Analysis (EDA)

1.Correlation Heatmap:

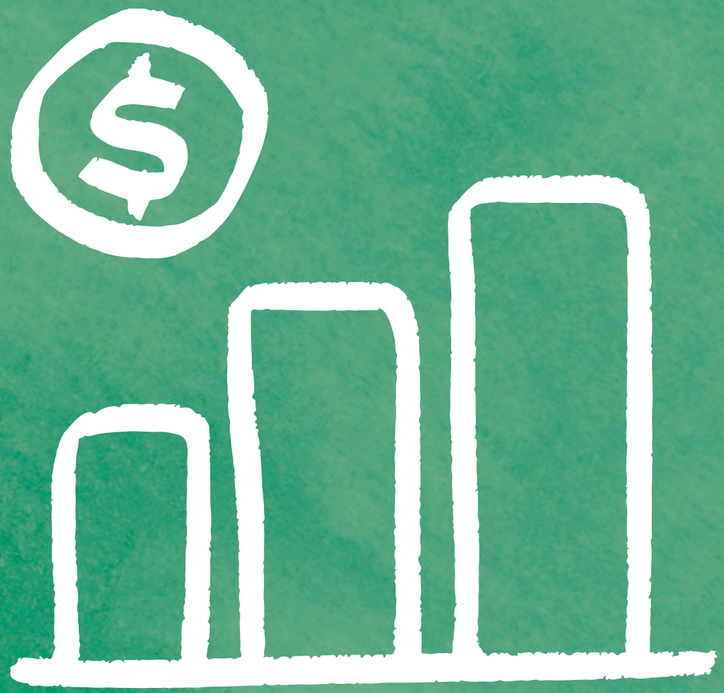
- Key insights: CTC is most correlated with role, experience, and previous CTC.

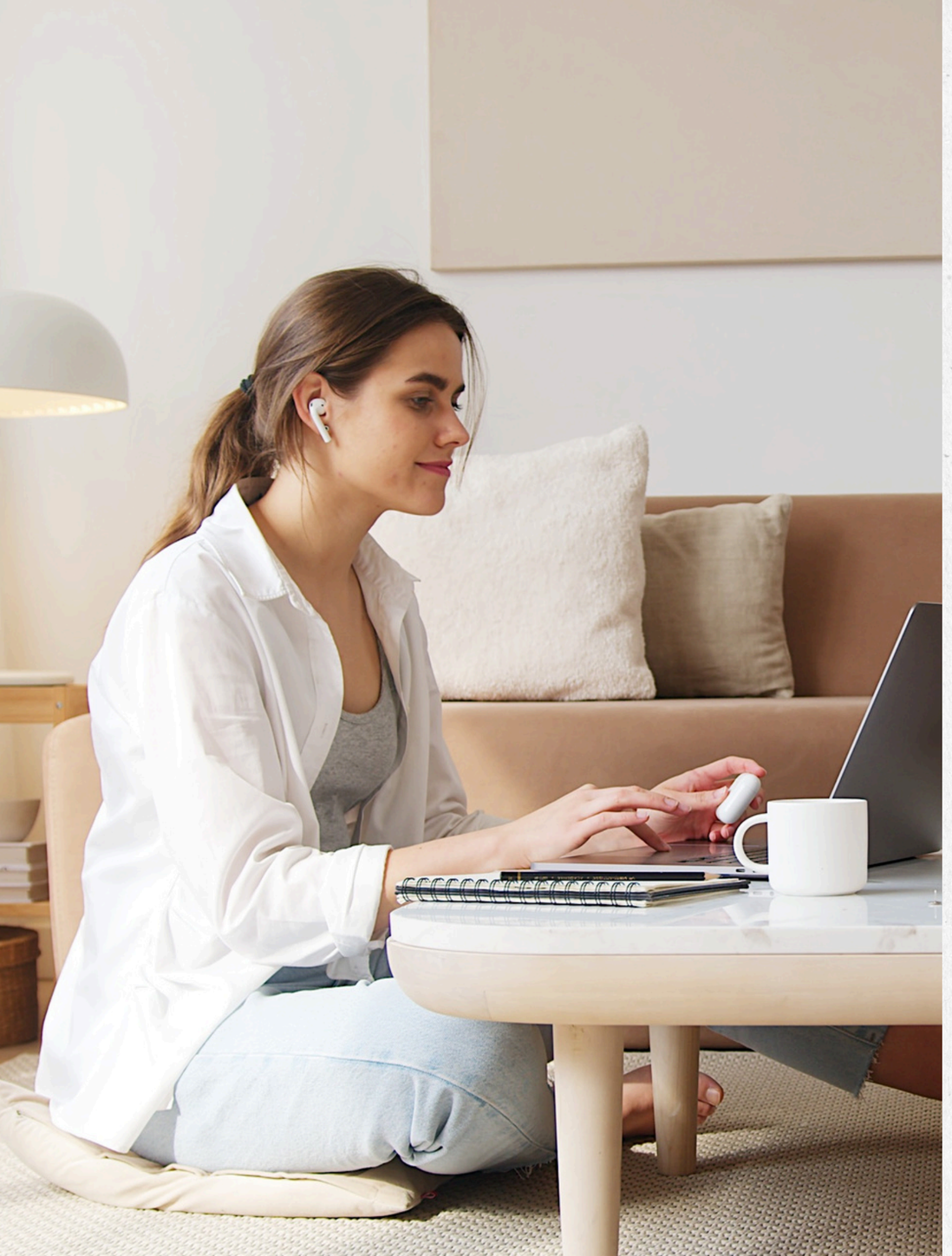
2.Distributions:

- Role, Experience (Months), Graduation Marks, and Previous CTC follow varied distributions.

3.Visualization Samples:

- Histograms of CTC, Graduation Marks, and Role.





Machine Learning Models

1. **Models Used:**
    - Linear Regression
    - Lasso and Ridge Regression
    - Decision Tree Regressor
    - Random Forest Regressor
    - Gradient Boosting Regressor
    - Bagging Regressor
    - XGBoost Regressor
2. **Evaluation Metrics:**
    - Mean Squared Error (MSE)
    - R^2 Score
    - Mean Absolute Percentage Error (MAPE)

Key Observations

- **Why Random Forest?**
  - Ensemble model that averages predictions from multiple trees, reducing overfitting.
  - Performs well on datasets with complex, non-linear relationships.
- **Other Models:**
  - Bagging and Gradient Boosting also performed well.
  - XGBoost underperformed due to possible lack of fine-tuning.

Conclusion

Summary:

- Predicting CTC is a complex task requiring thoughtful preprocessing, feature engineering, and model selection.
- Random Forest was the best-performing model, with an R^2 score of 60.14%

Thank you !