A logo of a company

Description automatically generated A close-up of a logo

Description automatically generated

**Joint Tech Internship Community Program**

**Assignment: Predicting Salary of Indian Software Engineers**

**Problem Statement:**

A tech recruitment agency wants to predict the annual salary of Indian software engineers based on various factors such as educational background, college tier, years of experience, academic performance, city of employment, skillsets, degree, certifications, and more. Your task is to develop a machine learning model that predicts the salary of software engineers given these attributes.

**Objective:**

Build a predictive model to estimate the annual salary of Indian software engineers. Evaluate the model using appropriate metrics and provide insights into the factors that most influence salary levels in the tech industry.

**Dataset:**

You are provided with a dataset containing the following columns:

1. **EngineerID**: Unique identifier for each software engineer.
2. **Age**: Age of the software engineer.
3. **Gender**: Gender of the software engineer (Male/Female).
4. **YearsOfExperience**: The number of years the engineer has been working in the software industry.
5. **Degree**: The highest degree attained by the engineer (e.g., Bachelor, Master, PhD).
6. **CollegeTier**: The tier of the college from which the engineer graduated (e.g., Tier 1, Tier 2, Tier 3).
7. **AcademicScore**: Average academic score or GPA of the engineer during their degree.
8. **City**: The city where the engineer is employed (e.g., Bangalore, Hyderabad, Pune, Chennai, Delhi).
9. **Skillsets**: The primary technical skillsets possessed by the engineer (e.g., Java, Python, Data Science, Cloud Computing).
10. **Certifications**: Number of professional certifications attained by the engineer.
11. **CompanySize**: The size of the company where the engineer is employed (e.g., Startup, Mid-size, Large Corporation).
12. **IndustryDomain**: The domain or industry in which the engineer is working (e.g., Finance, Healthcare, IT Services, E-commerce).
13. **WorkHoursPerWeek**: The average number of hours the engineer works per week.
14. **Salary**: The target variable representing the annual salary of the engineer (in INR).

**Tasks:**

1. **Data Exploration and Preprocessing:**
   * Load the dataset and perform initial exploration to understand the data.
   * Identify and handle any missing values appropriately.
   * Convert categorical variables into numerical ones using techniques such as One-Hot Encoding or Label Encoding.
2. **Feature Engineering:**
   * Perform feature scaling (e.g., Standardization or Normalization) on continuous variables.
   * Create new features if relevant, such as experience in relevant technologies or impact of certifications on salary.
   * Use dimensionality reduction techniques (e.g., PCA) to reduce the feature space if necessary.
3. **Model Building:**
   * Split the dataset into training and testing sets (e.g., 80-20 split).
   * Train different regression models (e.g., Linear Regression, Decision Trees, Random Forest, Gradient Boosting).
   * Perform hyperparameter tuning using techniques like Grid Search or Random Search to optimize model performance.
4. **Model Evaluation:**
   * Evaluate your models using appropriate metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.
   * Compare the performance of different models and select the best one.
   * Analyze feature importance to understand the most significant factors contributing to software engineer salaries.
5. **Insights and Recommendations:**
   * Provide insights based on your model analysis regarding the factors that influence software engineer salaries.
   * Suggest actionable strategies for engineers to optimize their career paths based on the findings.
6. **Documentation:**
   * Document your process, including data exploration, preprocessing steps, model selection, and evaluation.
   * Include visualizations where necessary to support your findings.

**Deliverables:**

* A Jupyter notebook (or Python script) with the entire workflow.
* A report summarizing your findings, including the model's performance and recommendations for optimizing salary expectations.

This assignment problem focuses on predicting the salary of Indian software engineers, a linear regression problem in the tech recruitment domain. The problem involves data preprocessing, feature engineering, model training, and evaluation, with a focus on understanding the factors that drive salary levels in the Indian tech industry.

Top of Form

Bottom of Form