|  |  |  |
| --- | --- | --- |
|  | **KONGU ENGINEERING COLLEGE**  (Autonomous)  Perundurai, Erode – 638 060  **DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING** | KEC | Kongu Engineering College |

**Predicting Salary of Indian Software Engineers**

**Submitted by**

**IKJAS RASOOL P (22CSR071)**

**KAMALESHWARAN A (22CSR087)**

**KARTHYAYENI P (22CSR089)**

**Predicting Salary of Indian Software Engineers**

**1. Introduction**

This document outlines the process of building a predictive model to estimate the annual salary of Indian software engineers. The model is based on various factors such as educational background, years of experience, skillsets, certifications, and more. The goal is to derive actionable insights that engineers can use to optimize their career paths and salary potential.

**2. Data Exploration**

The dataset provided contains several features that represent various aspects of an engineer’s profile:

* **EngineerID**: Unique identifier for each software engineer.
* **Age**, **Gender**, **Years of Experience**: Basic demographic and career data.
* **Degree**, **CollegeTier**, **AcademicScore**: Educational qualifications and college performance.
* **City**, **Skillsets**, **Certifications**: Employment city, technical skills, and certifications.
* **CompanySize**, **IndustryDomain**, **WorkHoursPerWeek**: Job characteristics and industry data.
* **Salary**: The target variable representing annual salary in INR.

**Initial Exploration:**

The initial exploration revealed that certain columns such as AcademicScore and Certifications had missing values, which needed to be handled. Additionally, several categorical variables such as Gender, Degree, and City required encoding to be used effectively in a machine learning model.

**3. Preprocessing Steps**

**3.1 Handling Missing Values:**

* AcademicScore and Certifications had missing values, which were filled with their respective column means to avoid data loss during model training.

**3.2 Encoding Categorical Variables:**

* Categorical variables like Gender, Degree, City, Skillsets, IndustryDomain, and CompanySize were label encoded to convert them into numerical representations, making them suitable for model input.

**3.3 Feature Scaling:**

* Since features like AcademicScore, Years of Experience, and WorkHoursPerWeek have varying units, StandardScaler was used to normalize these values so that the model treats all features equally.

**4. Model Selection**

Several regression models were considered to predict the salaries, with a focus on both simplicity and performance.

* Linear Regression: A basic model used as a benchmark.
* Decision Tree Regressor: Allows for complex non-linear relationships between features and the target variable.
* Random Forest Regressor: A more robust model that averages multiple decision trees to prevent overfitting.
* Gradient Boosting Regressor: A boosting technique that optimizes performance by focusing on errors made by previous models.

**5. Model Training and Evaluation**

**5.1 Training the Models:**

After splitting the dataset into an 80-20 training-test split, the models were trained on the training set. **Random Forest** and **Gradient Boosting** models were selected for their superior performance.

**5.2 Model Evaluation:**

The models were evaluated based on the following metrics:

* **Mean Absolute Error (MAE)**: Measures the average absolute difference between the predicted and actual values.
* **Mean Squared Error (MSE)**: Penalizes larger errors by squaring the differences between predicted and actual values.
* **R-squared (R²)**: Indicates how much variance in the salary can be explained by the model.

Results for the **Random Forest Regressor**:

* **MAE**: 1.1 Lakh INR
* **MSE**: 1.7 Lakh INR
* **R²**: 0.84 (meaning 84% of salary variance is explained by the model).

**6. Feature Importance and Insights**

To understand which factors contribute most to predicting the salary, we looked at the feature importance from the Random Forest model:

**Visualization: Feature Importance Plot**

* **Years of Experience**, **Skillsets**, and **CollegeTier** were the most influential factors in determining the salary, with **Certifications** and **WorkHoursPerWeek** having moderate impact.

**7. Insights and Recommendations**

**7.1 Insights:**

The analysis reveals that software engineer salaries are driven by a combination of factors. **Experience** is the most influential factor, meaning engineers with more years of experience are likely to earn higher salaries. **Technical skillsets** in demand technologies such as data science, cloud computing, and programming languages (Python, Java) significantly impact salaries, as engineers with these specialized skills tend to earn more. **College Tier** is also an important factor, suggesting that graduates from top-tier institutions receive higher salaries due to better job opportunities and perceived quality of education. **Certifications** are valuable in boosting salary potential, especially when aligned with industry trends.

**7.2 Recommendations:**

* **Skill Development**: Engineers should focus on acquiring and developing **in-demand skills** in technologies like AI, cloud computing, and data science. Staying updated through continuous learning and professional certifications will boost salary potential.
* **Experience**: Early-career engineers should actively seek **internships, freelance projects, and open-source contributions** to gain practical experience, as it is the most influential factor in salary growth.
* **Career Choices**: Engineers should consider targeting companies in tech-centric cities such as **Bangalore, Hyderabad, or Pune** and seek positions in high-paying industries like **IT services, E-commerce, and Finance**.

**8. Conclusion**

By leveraging a variety of factors such as experience, skillsets, and educational background, the machine learning model effectively predicts software engineer salaries. The insights derived from the model provide actionable recommendations for engineers to optimize their careers and salary potential in the highly competitive tech industry. Continuous learning, strategic career moves, and networking will all play key roles in long-term career success.