**Salary Prediction System**

**Project Report**

**<Version 1.0>**

Industrial Training (EDS)

**BACHELOR OF TECHNOLOGY (CSE) DATA SCIENCE**

**in COLLABORATION WITH TCS-iON**

|  |  |
| --- | --- |
| PROJECT GUIDE:  **Dr. AJAY RASTOGI** | SUBMITTED BY:  **MANYA JAIN**  **TCA2265023** |

**NOVEMBER 2024**



**FACULTY OF ENGINEERING & COMPUTING SCIENCES**

**TEERTHANKER MAHAVEER UNIVERSITY, MORADABAD**

**ACKNOWLEDGEMENT**

I would like to express my heartfelt gratitude to everyone who supported me throughout the completion of this project. I am deeply indebted to my project guide for their valuable guidance, encouragement, and constructive feedback, which were crucial in bringing this project to fruition.

I am thankful to my institution and faculty members for providing me with the necessary resources and a platform to explore and apply my skills in the field of machine learning. Their unwavering support and expert advice helped me tackle challenges during the development process.

I extend my gratitude to my peers for their collaboration and discussions that contributed to enhancing the quality of this project. Finally, I am grateful to my family for their continuous encouragement and motivation throughout this journey.

I hope this project serves as a small step toward solving real-world problems and contributes to the field of predictive modelling.

Student Name: Manya Jain

Place: CCSIT, TMU  
Date: November 25, 2024

**DECLARATION**

I hereby declare that this Project Report titled **Salary Prediction System** submitted by me and approved by my project guide, the College of Computing Sciences and Information Technology (CCSIT), Teerthanker Mahaveer University, Moradabad, is a bonafide work undertaken by me and it is not submitted to any other University or Institution for the award of any degree diploma / certificate or published any time before.

|  |  |  |
| --- | --- | --- |
|  |  | |
|  |  |  |
|  |  |  |
| **Student Name:** | MANYA JAIN  TCA2265023 |  |
|  |  |  |
| **Project Guide: (External)** |  |  |
| **Project Guide: (Internal)** | DR. AJAY RASTOGI |  |

**Brief About the Company**

**TCS-iON**

TCS-iON, a strategic business unit of Tata Consultancy Services (TCS), is a leading provider of technology-enabled solutions tailored to educational institutions, examination boards, and businesses. With a strong focus on simplifying complex processes, TCS-iON empowers organizations to achieve efficiency and scalability. By leveraging the technological prowess and global presence of TCS, TCS-iON ensures that clients receive robust, reliable, and innovative solutions.

1. **Collaboration with the B.Tech Data Science Program**

TCS iON has partnered with the B.Tech Data Science program to provide industry-aligned learning experiences, practical exposure, and insights into real-world applications of data science. This collaboration is aimed at bridging the gap between academia and industry, equipping students with the skills required to thrive in a competitive landscape. Through this partnership, students gain access to advanced tools, cutting-edge technology, and mentorship from industry professionals, ensuring a strong foundation in data science principles and methodologies.

1. **Role in the Project**

For this specific project, titled **"Salary Prediction using Machine Learning Models,"** TCS-iON's collaboration facilitated the application of theoretical knowledge to a real-world problem. The project focuses on predicting current salary levels based on factors such as age, years of experience, and past salary. This work aligns with the data-driven decision-making processes that TCS iON champions in its own business operations.

The company provided critical resources, including:

1. **Technological Tools**: Platforms and environments necessary for implementing machine learning models.
2. **Mentorship**: Expert guidance to ensure the proper structuring and development of the project.
3. **Real-World Insight**: Industry-aligned objectives that emphasize the practical importance of accurate salary predictions.
4. **Vision and Values**

TCS iON is committed to transforming education and business ecosystems through digital transformation. Its vision is to create seamless, efficient, and scalable solutions while maintaining its core values of integrity, excellence, and innovation. This project is an embodiment of the company's mission to encourage skill development and foster a mindset of problem-solving among students and professionals alike.

By supporting such initiatives, TCS iON not only strengthens its role as a thought leader in the technology space but also ensures that students are well-prepared to contribute to industry advancements, particularly in data science and machine learning.

This collaboration underscores TCS iON’s dedication to creating value-driven educational experiences and its impact on shaping future-ready professionals.

Table of Contents

[1 Project Title 6](#_Toc31652427)

[2 Problem Statement 6](#_Toc31652428)

[3 Project Description 6](#_Toc31652429)

[3.1 Scope of the Work 6](#_Toc31652430)

[3.2 Project Modules 6](#_Toc31652431)

[3.3 Context Diagram (High Level) 6](#_Toc31652432)

[4 Implementation Methodology 6](#_Toc31652433)

[5 Technologies to be used 6](#_Toc31652434)

[5.1 Software Platform 6](#_Toc31652435)

[5.2 Hardware Platform 7](#_Toc31652436)

[5.3 Tools, if any 7](#_Toc31652437)

[6 Advantages of this Project 7](#_Toc31652438)

[7 Assumptions, if any 7](#_Toc31652439)

[8 Future Scope and further enhancement of the Project 7](#_Toc31652440)

[9 Project Repository Location 7](#_Toc31652441)

[10 Definitions, Acronyms, and Abbreviations 8](#_Toc31652442)

[11 Conclusion 8](#_Toc31652443)

[12 References 9](#_Toc31652444)

**Appendix**

**A: Data Flow Diagram (DFD)**

**B: Entity Relationship Diagram (ERD)**

**C: Use Case Diagram (UCD)**

**D: Data Dictionary (DD)**

**E: Screen Shots**

# Project Title

***Salary Prediction System using Regression and Random Forest Models***  
The project aims to predict employees' salaries based on features like age, years of experience, and current salary, enabling HR teams to make informed salary decisions.

# Problem Statement

In many organizations, determining fair compensation is challenging due to a lack of standardized tools for salary prediction. This project addresses the problem by using machine learning models to predict salaries based on employee features. It streamlines decision-making and ensures transparency in salary determination.

# Project Description

**3.1 Scope of the Work**

* **In-Scope:**
  + Predict salaries based on input features: age, years of experience, and current salary.
  + Train and evaluate machine learning models: Linear Regression and Random Forest Regressor.
  + Provide insights into model performance using metrics and visualizations.
  + Save the better-performing model for future use.
* **Out of Scope:**
  + Real-time API integration.
  + Deployment in a production environment.
  + Handling additional features such as job roles or industry-specific factors.

**3.2 Project Modules**

* **Data Preprocessing Module:** Handles cleaning and preparing the dataset.
* **Model Training Module:** Trains Linear Regression and Random Forest models.
* **Evaluation Module:** Evaluates models using metrics like MSE, MAE, and R².
* **Visualization Module:** Visualizes predictions and errors using scatterplots and histograms.
* **Model Saving Module:** Saves the better-performing model for future predictions.
* **Web Hosting with Streamlit:** A Streamlit web application was developed to allow users to interact with the machine learning models through an intuitive interface. Users can input features such as age, years of experience, and current salary to predict their expected salary.

**3.3 Context Diagram (High Level)**

* **Inputs:** Age, Years of Experience, Current Salary (from HR Data).
* **Process:** Preprocessing → Model Training → Evaluation → Prediction.
* **Outputs:** Predicted Salary and Evaluation Metrics.

# Implementation Methodology

**The project involves:**

* **Data Preparation:** Using pandas to clean and preprocess the dataset.
* **Model Implementation:** Training Linear Regression and Random Forest models using scikit-learn.
* **Evaluation:** Comparing model performance using MSE, MAE, and R² scores.
* **Visualization:** Using seaborn for scatterplots (True vs. Predicted) and error histograms.
* **Testing:** Verifying model predictions on test data and ensuring consistent performance.

**Visuals Placement:**

* Place scatterplots (True vs. Predicted) in the **Evaluation Module** section.
* Include error distribution plots in the **Advantages of this Project** section to highlight the robustness of the Random Forest model.

**Web hosting:**

After the models were trained and evaluated, a **Streamlit-based web application** was developed to deploy the machine learning models. This application enables end-users to interact with the models and view the predictions directly. The Streamlit application integrates the pre-trained Random Forest model to ensure accurate results. The workflow for deployment involved:

* Exporting the trained model using Joblib.
* Developing an interactive interface using Streamlit's widgets like sliders and text inputs.
* Hosting the application locally or on a cloud platform (e.g., Streamlit Cloud).

**Project Modules (Streamlit Component):**

The **Streamlit module** provides a user-friendly interface for inputting user data (age, experience, and current salary) and predicting the expected salary using the trained machine learning model. The interface includes:

1. **Sliders** for age and experience, allowing dynamic input.
2. **Text fields** for entering current salary values.
3. A **predict button**, which processes the input and displays the predicted salary.

# Technologies to be used

**5.1 Software Platform**

* **Front-end:** **Streamlit** for hosting the application and creating an interactive interface.
* **Back-end:** Python, scikit-learn, pandas, seaborn, joblib.

**5.2 Hardware Platform**

* **RAM:** 8 GB.
* **Processor:** Intel core i5 10th generation
* **Hard Disk:** Minimum 20 GB free space.
* **OS:** Windows 10 Home.
* **Editor:** Jupyter Notebook, VS Code.
* **Browser:** Chrome/Firefox for documentation purposes.

**5.3 Tools**

* **seaborn** (Visualization)
* **joblib** (Model saving)
* **scikit-learn** (Model building and evaluation).

# Advantages of this Project

* Enables HR teams to make data-driven salary decisions.
* Improves accuracy compared to traditional heuristic methods.
* Scalable with additional features and larger datasets.
* Provides easy integration into existing HR tools.

# Assumptions, if any

* Data provided is clean and does not have missing values.
* No external API or real-time prediction module is required for the current scope.

# Future Scope and further enhancement of the Project

* Include more features such as education level, job role, and location for better predictions.
* Integrate the system with a web-based interface for real-time salary predictions.
* Apply advanced algorithms such as XGBoost or neural networks for further accuracy improvements.

# Project Repository Location

| **S#** | **Project Artifacts (softcopy)** | **Location** | **Verified by Project Guide** |
| --- | --- | --- | --- |
|  | Project Report (Final Version) | GitHub (https://github.com/Manya14Jain/TCS-ion-HRSalaryPrediction/blob/main/ManyaJainTCS-iON\_Report.docx) |  |
|  | Test Repository | GitHub (https://github.com/Manya14Jain/TCS-ion-HRSalaryPrediction) |  |
|  | Presentation | GitHub (https://github.com/Manya14Jain/TCS-ion-HRSalaryPrediction/blob/main/ManyaJainTCS-iON\_ppt.pptx) |  |

# Definitions, Acronyms, and Abbreviations

|  |  |
| --- | --- |
| **Abbreviation** | **Description** |
| MSE | Mean Squared Error |
| MAE | Mean Absolute Error |
| R² | Coefficient of Determination (Goodness of Fit) |

# Conclusion

The Salary Prediction System successfully demonstrates the use of machine learning for salary determination. Random Forest proved more accurate than Linear Regression, as shown by the lower MSE and higher R².

These models, combined with effective visualizations, provide actionable insights for HR teams. Future enhancements can make the system more robust and capable of handling dynamic business needs.

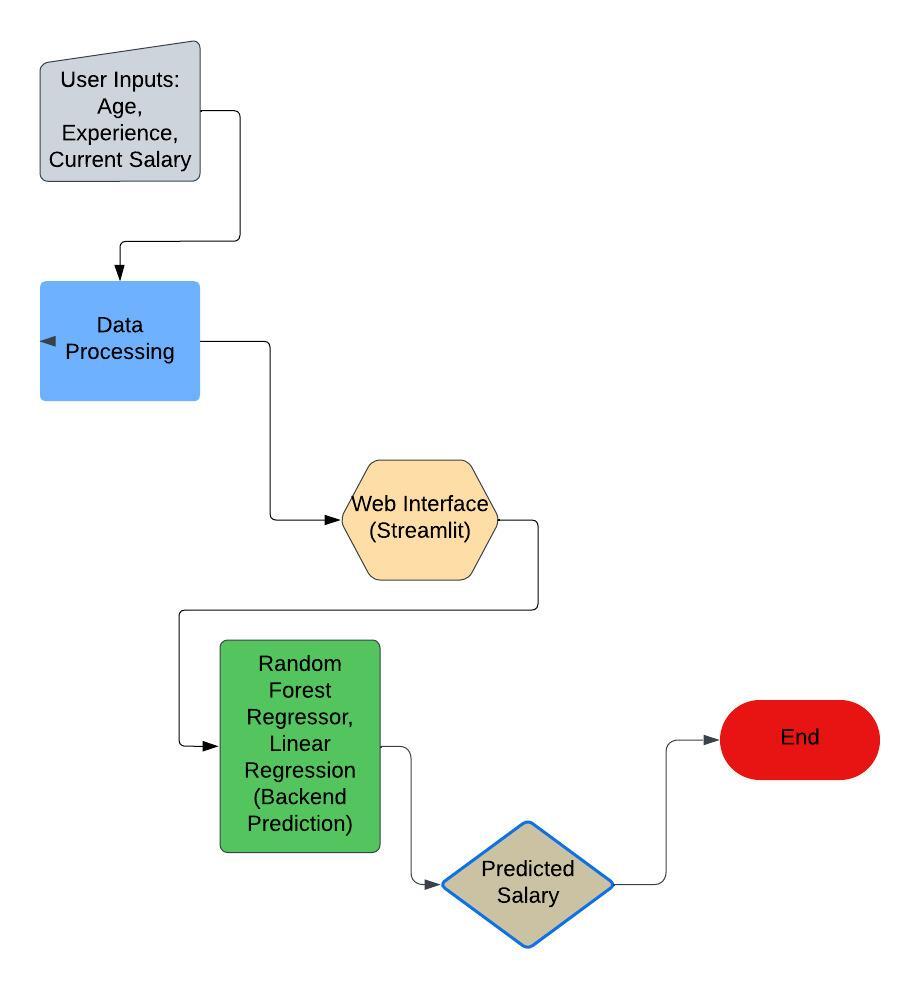
# References

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.no.** | **Reference Details** | **Owner** | **Version** | **Date** |
|  | Project Synopsis (GitHub) | Manya Jain | 1.0 | 25-11-24 |
|  | Project Requirements (GitHub) | Manya Jain | 1.0 | 25-11-24 |

**Annexure A**

**Data Flow Diagram (DFD)**

**(Mandatory)**

****

This diagram breaks down the system into more granular components:

1. **Input Data**: The user enters their details (Age, Years of Experience, and Current Salary) through the Streamlit interface.
2. **Model Prediction**: The input data is sent to the back-end where the Random Forest model processes it.
3. **Output**: The predicted salary is displayed on the web interface for the user.

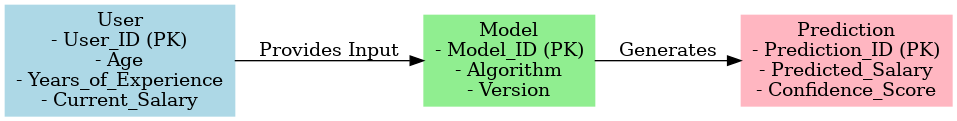
Here’s a step-by-step breakdown:

1. **Input**:
   * Age
   * Years of Experience
   * Current Salary
2. **Processing**:
   * The Streamlit application forwards inputs to the backend model.
   * Preprocessing is minimal, as inputs are directly formatted into a DataFrame compatible with the model.
   * The trained model predicts the salary based on the inputs.
3. **Output**:
   * Display the predicted salary on the web interface.

**Annexure B**

**Entity-Relationship Diagram (ERD)**

**(Mandatory)**

****

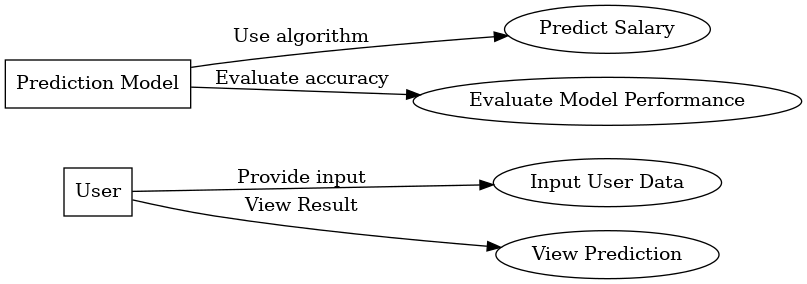
Entities and Attributes:

1. User
   * Attributes:
     + User\_ID: Unique identifier for the user (Primary Key).
     + Age: Age of the user.
     + Years\_of\_Experience: Total years of professional experience.
     + Current\_Salary: The current salary provided by the user.
2. Model
   * Attributes:
     + Model\_ID: Unique identifier for the model (Primary Key).
     + Algorithm: Name of the machine learning algorithm used (e.g., Random Forest).
     + Version: Version of the model.
3. Prediction
   * Attributes:
     + Prediction\_ID: Unique identifier for the prediction (Primary Key).
     + Predicted\_Salary: Output generated by the model.
     + Confidence\_Score: Measure of the model's confidence in the prediction.
4. Relationships:
5. User to Model:
   * Relationship: "Provides Input"
   * Description: The User submits their details (age, years of experience, current salary) to the Model for processing.
6. Model to Prediction:
   * Relationship: "Generates"
   * Description: The Model processes the input provided by the user and generates the Prediction, which includes the predicted salary and a confidence score.

**Annexure C**

**Use-Case Diagram (UCD)**

**(Optional)**

****

Actors:

1. User: Provides input data (age, years of experience, salary) and views the predicted output.
2. Prediction Model: Represents the ML models (Linear Regression and Random Forest Regressor) that process input data.

Use Cases:

1. Input Data: User submits input data through the web interface.
2. Predict Salary: ML models compute the expected salary based on user inputs.
3. View Prediction: Predicted salary is displayed along with evaluation metrics and scatter plots.
4. Evaluate Performance: Application calculates metrics (MSE, R²) for both models.

Relationships:

* The User interacts with the system to provide input and view predictions.
* The Prediction Model processes the input data, performs predictions, and computes metrics.

**Diagram:** A UCD visually shows:

* **Actors** (stick figures): User and Prediction Model.
* **Use Cases** (ellipses): Input Data, Predict Salary, View Prediction, Evaluate Performance.
* **Relationships** (lines/arrows): Link actors to use cases.

**Annexure D**

**Data Dictionary (DD)**

**(Mandatory)**

**User Table (USER)**

|  |  |  |
| --- | --- | --- |
| **Field** | **Data Type** | **Description** |
| USER\_ID | Integer | Unique identifier for each user |
| AGE | Integer | Age of the user |
| YEARS\_OF\_EXPERIENCE | Integer | Years of experience in the current field |
| CURRENT\_SALARY | Float | Current salary of the user |

**Model Table (MODEL)**

|  |  |  |
| --- | --- | --- |
| **Field** | **Data Type** | **Description** |
| MODEL\_ID | Integer | Unique identifier for each model |
| ALGORITHM | Text | Algorithm used (e.g., Random Forest, Linear Regression) |
| VERSION | Text | Version of the model |

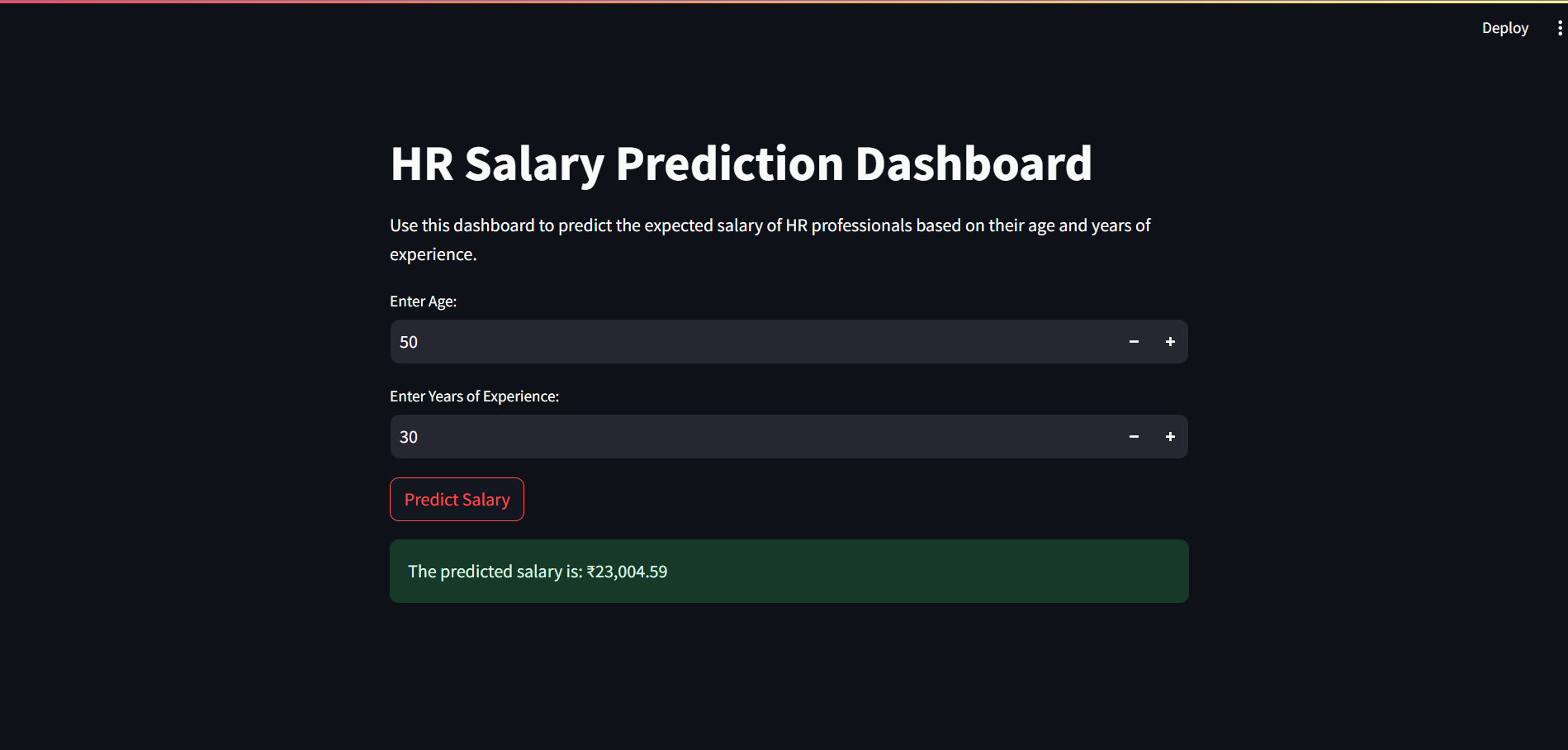
**Prediction Table (PREDICTION)**

|  |  |  |
| --- | --- | --- |
| **Field** | **Data Type** | **Description** |
| PREDICTION\_ID | Integer | |  | | --- | |  |  |  | | --- | | Unique identifier for each prediction | |
| PREDICTED\_SALARY | Float | Predicted salary value |
| CONFIDENCE\_SCORE | Float | |  | | --- | |  |  |  | | --- | | Confidence score of the prediction | |
| USER\_ID | Integer | Reference to the user table (foreign key) |
| MODEL\_ID | Integer | Reference to the model table (foreign key) |

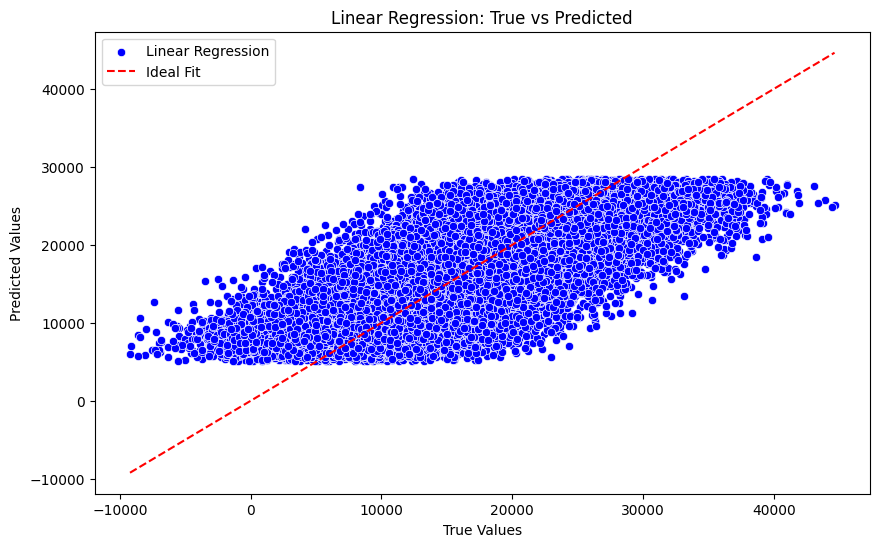
**Annexure E**

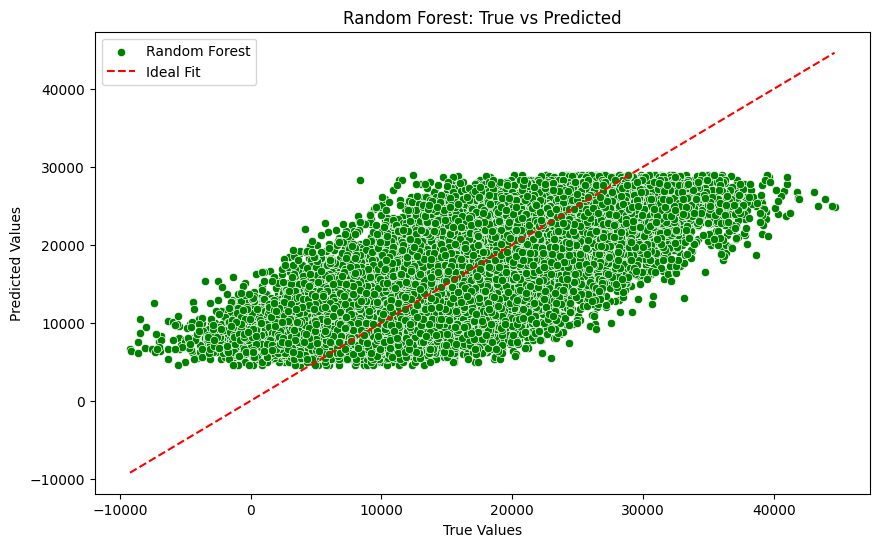
**Screen Shots**

**Home Page:**



**Model Visuals:**

****

**

**Model Metrics:**

