Salary PREDICTION

By Sanhita Saxena

Project INTRODUCTION

This project predicts employee salaries using advanced regression techniques and thorough data preprocessing. We compare multiple machine learning models to select the best-performing one, which is then integrated into a pipeline. This pipeline includes all necessary preprocessing steps, ensuring accurate predictions and easy deployment.

Project OBJECTIVES

01 EXPLORATORY DATA ANALYSIS

02 DATA PREPROCESSING

03 FEATURE ENGINEERING

04 FEATURE SELECTION

05 MODEL TRAINING AND EVALUATION

06 MODEL COMPARISON AND SELECTION

07 HYPERPARAMETER TUNING

08 PIPELINE BUILDING

09 DEPLOYMENT

Exploratory DATA ANALYSIS

- **Gender Distribution**: More females than males.
- **Unit Distribution**: Most employees are in the IT department.
- **Designation Distribution**: A significantly large population of analysts.
- **Age Trend**: Older employees tend to have higher salaries.
- **Experience Trend**: Higher experience correlates with higher salaries.

Data PREPROCESSING

- **Dropped** duplicates.
- **Checked** for null values.
- **Imputed** DOJ, AGE, and RATINGS columns with mode.
- **Imputed** LEAVES USED, LEAVES REMAINING columns with median.
- **Dropped** FIRST NAME, LAST NAME columns.

Feature ENGINEERING

- **Converted** DOJ and CURRENT DATE columns to datetime datatype.
- **Created** a new feature: years_experience.
- **Dropped** date columns.

Feature SELECTION

- **Used** SelectKBest to select best features.
- **Extracted** selected feature names: SEX, DESIGNATION, AGE, UNIT, PAST EXP, years_experience.
- **Reassigned** selected features to training and test datasets.
- **Conducted** correlation study and dropped columns with correlation exceeding [-0.8, 0.8].

Model TRAINING *and* EVALUATION

- **Developed** helper functions for calculating and storing regression metrics.
- **Implemented** functions to train models and generate predictions.
- **Evaluated** model performance by comparing training and test metrics.

Model COMPARISON and SELECTION

- **Compared** multiple regression models, including Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, XGBoost, and Support Vector Regression.
- **Evaluated** each model using Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and R2 score.
- **Selected** Gradient Boosting Regression as the best-performing model with an R2 score of 0.9495 for final pipeline integration.

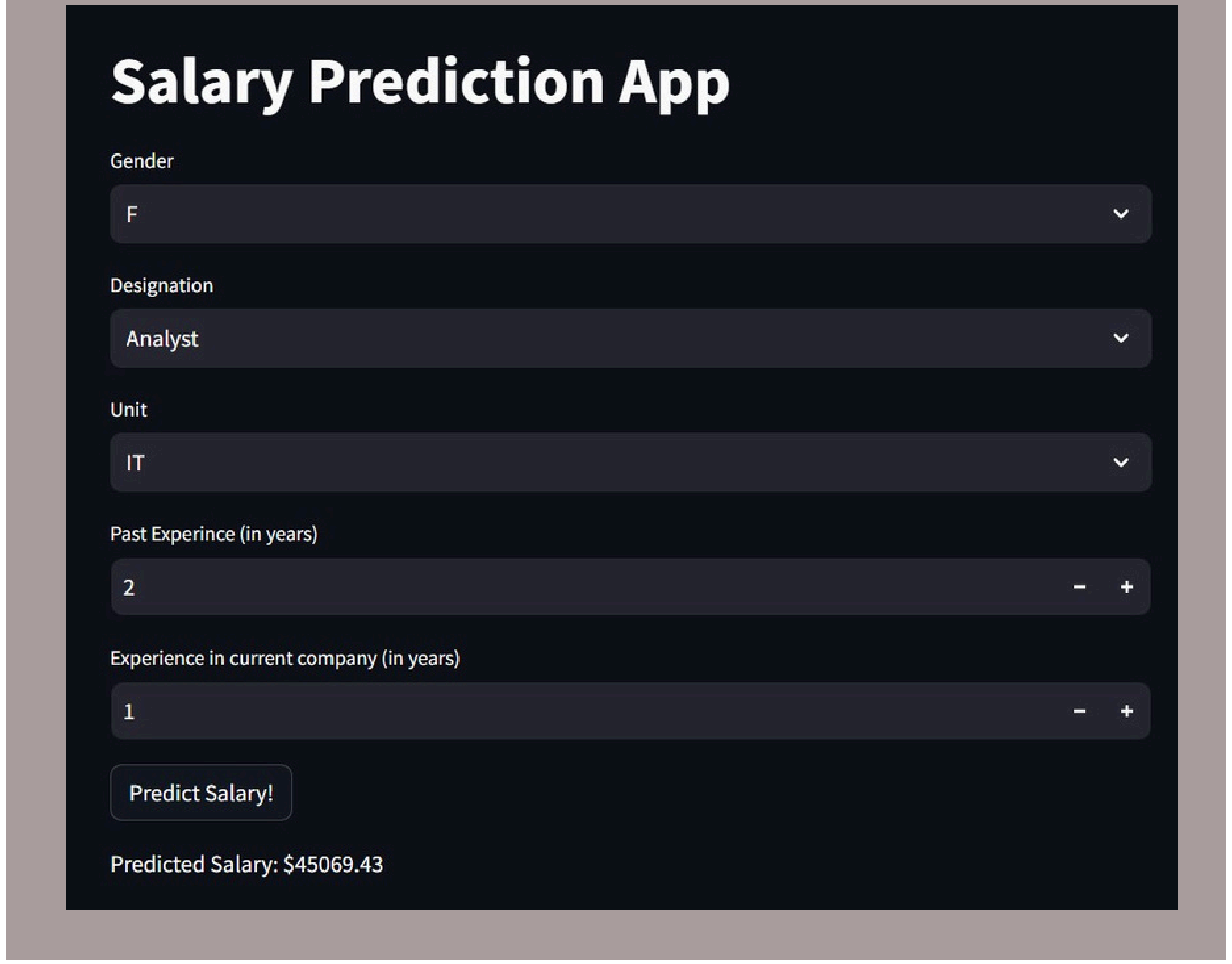
Hyperparameter TUNING

- **Created** a pipeline that included data preprocessing steps, feature selection with SelectKBest, and the Gradient Boosting Regressor model.
- **Conducted** hyperparameter tuning using RandomizedSearchCV to optimize the Gradient Boosting Regressor.
- **Tuned** hyperparameters such as `n_estimators`, `learning_rate`, `max_depth`, `min_samples_split`, `min_samples_leaf`, and `subsample`.
- **Achieved** improved model performance by identifying the best combination of hyperparameters for the Gradient Boosting Regressor.

Pipeline BUILDING

- Built a preprocessing pipeline using ColumnTransformer to handle different types of categorical encoding (ordinal and nominal).
- Integrated SelectKBest for feature selection within the pipeline to automatically choose the most relevant features.
- Incorporated the best-performing model, Gradient Boosting Regressor, into the pipeline.
- Configured the final pipeline to include all preprocessing steps, feature selection, and the regression model for seamless integration and deployment.
- Ensured that the entire pipeline, including preprocessing and model training, could be saved and reused efficiently.

Web App DEPLOYMENT



Salary Prediction App

Gender: F

Designation: Analyst

Unit: IT

Past Experince (in years): 2

Experience in current company (in years): 1

Predicted Salary: $45069.43

- **Developed** a web application using Streamlit for easy and interactive salary prediction.
- **Created** a user-friendly interface for inputting employee information and obtaining salary predictions.
- **Deployed** the application using Streamlit's built-in sharing platform.
- **Included** a `requirements.txt` file to ensure easy replication of the development environment for deployment.
- Check it out here.

Thank You!

Email | Linkedin | GitHub