Assignment: Data Science Salaries Prediction Model with MLflow and Streamlit

Create a machine learning model to predict data science salaries using the dataset provided. Utilize MLflow for tracking and managing the model training process. Develop a user interface using Streamlit to interact with the model. Finally, register and deploy the best model using Hugging Face Spaces.

Dataset:

Use the Data Science Salaries 2023 dataset available on Kaggle: Data Science Salaries 2023.

Tasks and Requirements:

1. **Data Exploration and Preprocessing:**
   - Load the dataset and perform exploratory data analysis (EDA).
   - Clean the data, handle missing values, and encode categorical variables.
   - Split the data into training and testing sets.
2. **Model Training:**
   - Train multiple machine learning models (e.g., Linear Regression, Decision Trees, Random Forest, Gradient Boosting).
   - Use MLflow to track experiments, including parameters, metrics, and artifacts.
   - Evaluate the models using appropriate metrics (e.g., RMSE, MAE, R^2).
3. **Model Selection and Optimization:**
   - Compare the performance of different models.
   - Optimize the best-performing model using hyperparameter tuning.
   - Record all experiments and their results using MLflow.
4. **Streamlit Application:**
   - Create a Streamlit app to interact with the trained model.
   - The app should allow users to input features and get salary predictions.
   - Display relevant model performance metrics and visualizations in the app.
5. **Model Registration and Deployment:**
   - Register the best model in the MLflow Model Registry.
   - Deploy the model using Hugging Face Spaces.
   - Ensure the deployed model is accessible via an API for inference.

Deliverables:

Data loading and EDA steps.

Data preprocessing steps.

Model training and evaluation steps.

Hyperparameter tuning steps.

MLflow tracking code.

A Streamlit app directory with all necessary files and a `requirements.txt` file.

The app should be hosted and accessible via a public URL.

The registered model in the MLflow Model Registry.

The deployed model on Hugging Face Spaces with a public API endpoint.

Evaluation Criteria:

- **Completeness:** All tasks and deliverables are completed as specified.
- **Code Quality:** The code is well-documented, organized, and follows best practices.
- **Innovation:** Creativity in model selection, feature engineering, and application design.
- **Presentation:** Clarity and usability of the Streamlit application.
- **Deployment:** Successful deployment and accessibility of the model via Hugging Face Spaces.