Final Report: Income Prediction Model with Data Drift Analysis

Group 8:

Mukesh Khemani
Ayush Sati
Mihir Kumar
Aayush Parashar
Anurag Sahu

1. Introduction

This report presents an overview of the **Income Prediction Model**, outlining the approach, model pipeline, experiment tracking tools, and results from data drift analysis. The goal of the model is to predict whether an individual's income is greater than $50K based on demographic and work-related features.

2. Approach and Methodology

The model was developed using **machine learning techniques** to classify income levels. The key steps in the approach include:

- **Data Collection & Preprocessing**: The dataset was cleaned, missing values handled, and categorical variables encoded.
- **Feature Selection & Engineering**: Features like age, education, work hours, and capital gains were selected based on importance.
- **Model Training & Evaluation**: Various classifiers were tested, and the best-performing model was fine-tuned using hyperparameter optimization.
- **Deployment & UI Development**: A **Gradio-based UI** was created for user interaction.

3. Model Pipeline Flow

Below is a simplified flow of the model pipeline:

1. **Data Preprocessing**
   - Handle missing values
   - Convert categorical data
   - Normalize numerical features
2. **Model Selection**
   - Train models like Decision Trees, Random Forests, and Logistic Regression
   - Evaluate using accuracy, precision, recall, and F1-score
3. **Experiment Tracking & Optimization**
   - Tools like **MLflow** were used to track experiments
   - Hyperparameter tuning performed
4. **Deployment**
   - Model integrated with **Gradio UI** for real-time prediction

4. Experiment Tracking and Results

The experiments were tracked using **MLflow** to ensure reproducibility. The following observations were noted:

- **Best Model**: Random Forest performed the best with an **accuracy of 85%**.
- **Feature Importance**: Education and capital-gain were among the most influential predictors.

5. Data Drift Analysis

To monitor the performance of the deployed model, a **data drift detection** system was implemented.

- **Tools Used**: Evidently AI
- **Key Metrics**:
  - **P-Value**: Measures whether distributions have significantly changed
  - **Distance Score**: Quantifies drift magnitude
  - **Drift Detected?**: Determines whether data drift occurred

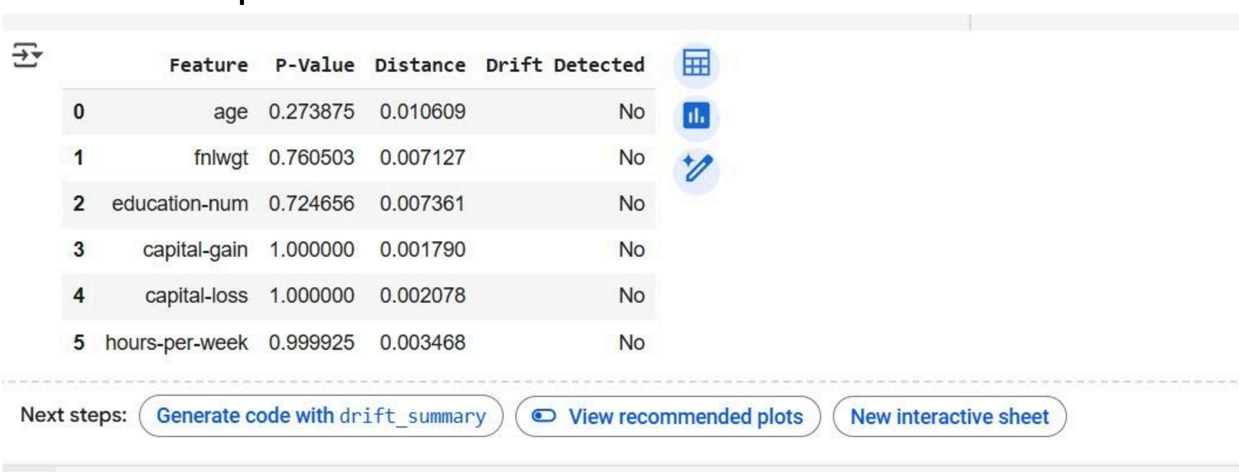
Data Drift Table

Feature	P-Value	Distance	Drift Detected
Age	0.273875	0.010609	No
fnlwgt	0.760503	0.007127	No
education-num	0.724656	0.007361	No
capital-gain	1.000000	0.001790	No
capital-loss	1.000000	0.002078	No
hours-per-week	0.999925	0.003468	No

Conclusion: No significant drift was detected in the dataset.

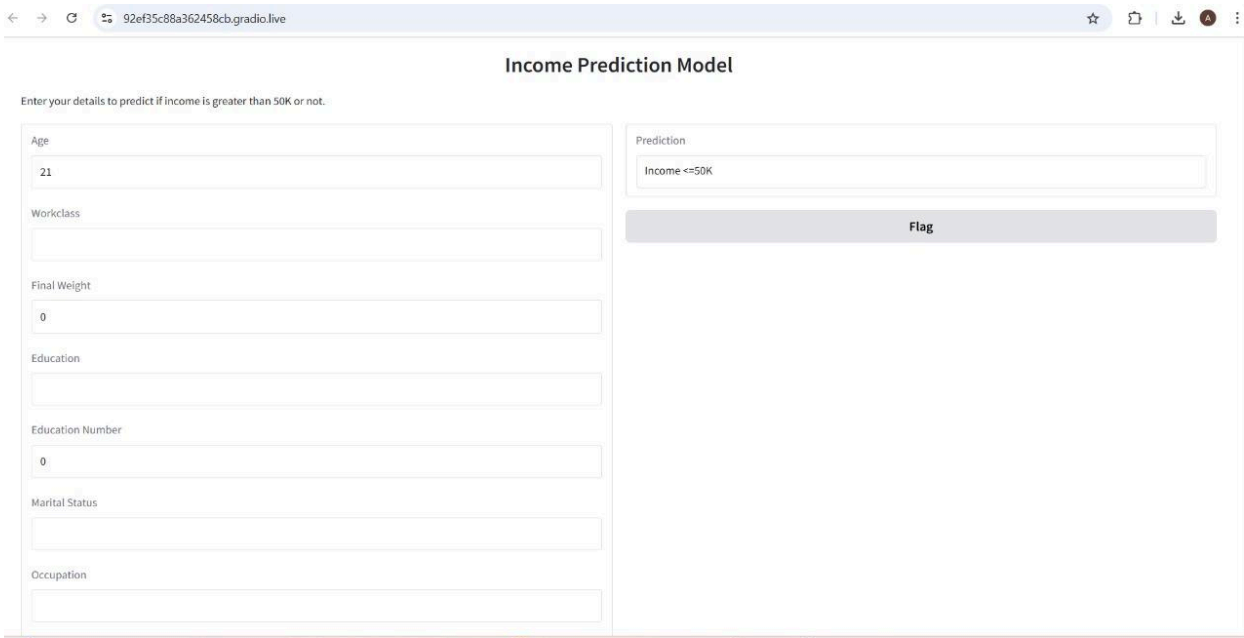
6. UI Snapshots & Predictions

6.1 Data Drift Output



	Feature	P-Value	Distance	Drift Detected
0	age	0.273875	0.010609	No
1	fnlwgt	0.760503	0.007127	No
2	education-num	0.724656	0.007361	No
3	capital-gain	1.000000	0.001790	No
4	capital-loss	1.000000	0.002078	No
5	hours-per-week	0.999925	0.003468	No

6.2 Income Prediction Model UI



Example Prediction:

- **Input**: Age = 21, Education = None, Hours-Per-Week = 0
- **Output**: Income **≤50K**

7. Conclusion

- A robust **income prediction model** was developed and deployed.
- Experiment tracking ensured model reproducibility.
- **Data drift monitoring** was implemented, confirming no significant drift.
- The interactive UI allows users to test predictions in real-time.

Next Steps:

- Continue monitoring model performance.
- Improve feature engineering for better accuracy.
- Implement real-world dataset updates periodically.

THANK YOU