Project Title: Identifying Data Integration Quality for Multi – Source

Name: Chethana N

CAN_ID: 33760732

Institution Name: Vemana Institute of Technology

1. Overview of Results

This phase focuses on evaluating the model's performance in identifying data integration quality. The performance metrics and visualizations provide insights into how well each model handles data quality issues like anomalies, missing values, and bias. A dashboard was developed to enhance user interaction by visualizing flagged records and tracking quality scores dynamically.

---

2. Performance Metrics and Visualizations

The evaluation metrics for different models are as follows:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	92.3%	88.5%	91.2%	89.8%
Random Forest	95.7%	94.0%	95.1%	94.5%
AutoAI Model	97.1%	96.5%	96.8%	96.6%

Insights:

- The AutoAI model showed the best performance, achieving high accuracy, precision, and recall with a balanced F1-Score.
- **Confusion Matrix**: Showed high True Positives (TP), low False Positives (FP), and moderate False Negatives (FN), highlighting the model's strong detection capability with minimal errors.
- **ROC Curve**: Demonstrated robust classification, indicating excellent separation between quality and non-quality data across all models.

---

3. Dashboard Development

A dashboard was developed using **HTML, CSS, JavaScript, and Bootstrap** to provide an interactive way of visualizing flagged transactions.
Key Features:

1. **Upload Data**: Users can input data arrays to be analyzed.
2. **Flagged Records**: Displays records that the model identifies as low-quality.
3. **Visual Charts**: Dynamic visualizations to understand patterns in flagged data.

This dashboard improves transparency and provides stakeholders with actionable insights.

---

4. Model Deployment on IBM Cloud

The AutoAI model was deployed on **IBM Cloud** using the following steps:

1. **Create IBM Cloud Account:** Signed up for IBM Cloud and enabled Watson Studio and AutoAI services.
2. **Upload Model:** Uploaded the trained AutoAI model and deployed it as a REST API.
3. **Generate API Key:** Generated an API key for secure interaction with the deployed model.**Server Development:** Created a Node.js server to interface with the deployed odel and fetch predictions for new data inputs.

This setup allows seamless integration of the model with the web-based dashboard for real-time analysis and predictions.

---

5. Resource Usage and Scalability

The project utilized the following resources on IBM Cloud:

- **Compute Hours (CUH):** Approximately 14 CUH used for data preprocessing, model training, and deployment.
- **API Requests:** Limited to 50 requests per month under the Lite plan, sufficient for testing purposes.
- **Storage:** 1 GB of free storage for datasets and models.
  To scale, options like transitioning to paid plans and optimizing workflows can help accommodate larger datasets and more frequent API calls.

---

6. Future Enhancements

To further enhance the project:

1. **Deploy UI on Vercel:** Make the dashboard accessible to end users by deploying it on Vercel.
2. **CSV Upload Feature:** Allow users to upload CSV files for more flexible data analysis.
3. **Database Integration:** Store records for historical analysis and improve scalability.
4. **Automated Reports:** Generate detailed reports on detected anomalies and system performance.

---

7. Conclusion

The project successfully integrated advanced data quality detection models with a user-friendly web interface and smooth deployment on IBM Cloud. The AutoAI model's superior performance highlights the potential for automating data quality detection with minimal manual intervention. By balancing efficiency, scalability, and user experience, the project lays a solid foundation for future improvements in data integration quality analysis.