**User Story: Exploring Sales Data and Model Evaluation**

**As** a Data Scientist at X Company, **I want** to analyse and visualize our sales data for the year 2022, compare it to the actual sales data from 2021, and evaluate the performance of a predictive model we built for customer type classification. **So that** I get benefits to understand the customer profile in region wise, month wise and product category wise. Helps me to know the staff demand as well.

**Acceptance Criteria:**

1. **Data Loading and Model Training:**

- Given a dataset named 'Company\_2021.csv' containing sales data for 2021 and 'Company\_2022.csv' for 2022,

- And a trained on multiple classifiers like, Random Forest (n\_estimator=100), Decision tree, SVM (Kernel = linear), K-Nearest Neighbors (n\_neighbors = 5), Logistic Regression, Gradient Boosting (n\_estimators = 100)

- When I run the provided code,

- Then the data should be loaded, features and target variable defined, and the model trained on 2021 data.

2. **Predictions and Model Evaluation:**

- Given predictions made by the model on the 2022 data,

- And the original 'Company\_type\_original.csv' data containing actual customer types for 2022,

- When I compare the predicted and actual customer types,

- Then the accuracy of the model on the 2022 data should be displayed.

- And a classification report should be generated, showing precision, recall, and F1-score for each customer type.

3. **Visualization of Results:**

- Given the actual and predicted customer types for 2022,

- When I run the code to create subplots for 'monthnum,' 'region,' and 'product',

- Then three subplots should be displayed, comparing actual vs. predicted customer types by month, region, and product.

- And for each subplot, the x-axis should represent the corresponding category, and the y-axis should represent the customer type.

4. **Staff Quantity and Cost Analysis:**

- Given a dataset named 'Company.xlsx' containing staff quantity and cost data for 2021 and 2022,

- When I run the code to create subplots for staff quantity and staff cost by month for both years,

- Then four subplots should be displayed, showing the trends of staff quantity and cost for 2021 and 2022 by month.

- And each subplot should have the month on the x-axis and either staff quantity or staff cost on the y-axis.

5. **Save and Share:**

- When I want to save the generated plots,

- Then I can use the provided code to save the plots as image files for future reference or sharing with colleagues.

**Definition of Done:**

- The code should run without errors and produce accurate visualizations.

- Model evaluation metrics and visualizations should help assess the model's performance and trends in staff quantity and cost.

- The saved images should be named appropriately for easy reference.

---

This user story outlines the objectives, criteria for success, and expected outcomes of using the provided code to analyse and visualize sales data and model performance.