Report 4 phase

Assigned	Ninel Yunusova
Projects	Phase IV - Model validation and deployment
Status	Done

Chapter 5: Model Evaluation

Model training is followed by a crucial model evaluation phase, also known as offline testing. In this phase, the performance of the trained model is validated on a test set, and its robustness is assessed with noisy or erroneous input data. It is also essential to develop an explainable ML model to foster trust, meet regulatory requirements, and guide humans in ML-assisted decisions. The decision to deploy the model should be made based on predefined success criteria, with input from both domain and ML experts. All outcomes of the evaluation phase must be thoroughly documented.

Section 5.1: Model Validation Report

Performance on the Test Dataset:

The model's performance on the test dataset was assessed using the Mean Absolute Error (MAE) and Mean Squared Error (MSE) metrics. The results indicate that the Decision Tree Regressor achieved a satisfactory performance on the test set.

- **MAE**: 10.5
- **MSE**: 1209.193

Vulnerabilities Identified by Giskard:

The Giskard tool identified several vulnerabilities in the model, highlighting areas where the model's performance significantly deviated from the global metrics:

1. `clean_description_col_vector_6` **< 0.217**:
    - **MSE**: 4088.309 (238.10% higher than global)
    - **Samples Affected**: 140 (5.6% of dataset)

2. `clean_description_col_vector_7` **< 0.055 AND** `clean_description_col_vector_7` **>= 0.002**:
    - **MSE**: 3802.618 (214.48% higher than global)
    - **Samples Affected**: 128 (5.1% of dataset)
3. `clean_description_col_vector_3` **< 0.095 AND** `clean_description_col_vector_3` **>= 0.084**:
    - **MSE**: 3600.250 (197.74% higher than global)
    - **Samples Affected**: 127 (5.1% of dataset)
4. `San Francisco` **== 1.000**:
    - **MSE**: 2716.813 (124.68% higher than global)
    - **Samples Affected**: 547 (21.8% of dataset)
5. `Move_in_date_day` **< -7.176e-02 AND** `Move_in_date_day` **>= -1.866e-01**:
    - **MAE**: 18.186 (108.95% higher than global)
    - **Samples Affected**: 163 (6.5% of dataset)
6. `Day_of_the_week_recorded` **>= 3.500 AND** `Day_of_the_week_recorded` **< 4.500**:
    - **MSE**: 1835.789 (51.82% higher than global)
    - **Samples Affected**: 394 (15.7% of dataset)
7. `Beds` **>= -1.126e+00 AND** `Beds` **< 0.285**:
    - **MAE**: 12.601 (44.78% higher than global)
    - **Samples Affected**: 1273 (50.7% of dataset)
8. `Floor` **< -5.287e-01 AND** `Floor` **>= -6.648e-01**:
    - **MSE**: 1747.351 (44.51% higher than global)
    - **Samples Affected**: 495 (19.7% of dataset)
9. `Units` **>= -1.180e+00 AND** `Units` **< -8.622e-01**:
    - **MAE**: 12.418 (42.69% higher than global)
    - **Samples Affected**: 243 (9.7% of dataset)
10. `Boston` **== 1.000**:
    - **MSE**: 1710.406 (41.45% higher than global)

- **Samples Affected**: 357 (14.2% of dataset)

11. `Day_of_the_week_recorded` **>= 2.500 AND** `Day_of_the_week_recorded` **< 3.500**:
    - **MSE**: 1710.314 (41.44% higher than global)
    - **Samples Affected**: 388 (15.4% of dataset)
12. `clean_rn_col_vector_0` **< -1.671e+00 AND** `clean_rn_col_vector_0` **>= -2.047e+00**:
    - **MAE**: 11.282 (29.63% higher than global)
    - **Samples Affected**: 190 (7.6% of dataset)
13. `Northern_Exposure` **== 1.000**:
    - **MSE**: 1545.993 (27.85% higher than global)
    - **Samples Affected**: 515 (20.5% of dataset)
14. `Baths` **< -2.308e-01**:
    - **MSE**: 1538.018 (27.19% higher than global)
    - **Samples Affected**: 1597 (63.6% of dataset)
15. `Eastern_Exposure` **== 1.000**:
    - **MSE**: 1537.430 (27.15% higher than global)
    - **Samples Affected**: 559 (22.3% of dataset)

Section 5.2: Discussion

Comparison of ML Modeling and Giskard Validation:

The initial ML modeling using the Decision Tree Regressor produced a solid performance with an MAE of 10.5 and MSE of 1209.193 on the test dataset. However, the Giskard validation highlighted significant vulnerabilities in specific segments of the dataset where the model's performance degraded substantially. For instance, attributes like `clean_description_col_vector_6` and geographic locations such as `San Francisco` showed a considerable increase in error metrics, indicating areas where the model may be overfitting or struggling to generalize.

Comparison with Defined Success Criteria:

The defined success criteria for model deployment included achieving an MAE below 15 and ensuring robustness across various data segments. While the overall MAE on the test dataset met this criterion, the substantial variances and high error rates in specific segments identified by Giskard suggest that the model may not be sufficiently robust for deployment. These findings underscore the need for further refinement to address these vulnerabilities and enhance the model's generalization capabilities.

Section 5.3: Deployment Decision

Based on the comprehensive evaluation and discussion, the decision on whether to deploy the model involves weighing the model's overall performance against its identified weaknesses. Although the Decision Tree Regressor achieved an acceptable MAE on the test dataset, the significant performance degradation in certain segments highlighted by Giskard raises concerns about its robustness and reliability in production environments.

Decision:
At this stage, it is recommended
to deploy the model with identified vulnerabilities. We have reached the business success criteria and unfortunately do not have more time for more detailed development