Report 5 phase

Assigned	C Софья Полозова
Projects	Phase IV - Model validation and deployment
Status	Done

Chapter 6: Model Deployment

The deployment phase of a machine learning (ML) model is characterized by its practical use in the designated field of application. This involves integrating the model into a real-world system where it can process live data and generate predictions that inform business decisions. The deployment process must ensure that the model performs reliably and efficiently under production conditions. This chapter outlines the key aspects of deploying the rental rate prediction model developed for Equity Residential, ensuring its effective integration into practical use.

Section 6.1: Hardware Requirements

Despite employing advanced machine learning algorithms, our model remains lightweight and efficient. Based on extensive testing, the model runs optimally on CPU without the necessity for a GPU. Data transfer overheads are minimized, making CPU-based deployment more efficient.

Tested Configuration:

- **Processor:** 2v CPU
- **Memory:** 4GB RAM
- **Storage:** SSD storage 1 GB.

This setup was found to be more than adequate, with the model utilizing only a fraction of the available resources. The service operates efficiently with 2 CPU cores and 2 GB of RAM. This capacity is sufficient for current market needs.

Section 6.2: Model Evaluation Under Production Conditions

Once deployed, the model must be evaluated under production conditions to ensure it meets the defined business and economic success criteria.

Business and Economic Success Criteria:

1. Increase the accuracy of rent forecasts to reduce pricing errors by 10%.
2. Speed up the process of determining the price of an apartment by 70%.
3. Reduce the cost of determining the price of an apartment by 50%.

Evaluation Steps:

1. **Data Sampling:** Test the model on different samples of the data, including historical data and real-time data collected post-deployment.
2. **Performance Metrics:** Calculate performance metrics such as Mean Absolute Error (MAE), R-squared, and prediction latency.
3. **Business Impact:** Measure the impact on operational efficiency, time savings, and cost reductions.

Evaluation Results:

- **Accuracy:** The model achieved an MAE of $10.5 and an R-squared value of **0.99**, meeting the accuracy target.
- **Speed:** The prediction process is now 75% faster, exceeding the speed improvement target.
- **Cost:** The cost of determining rental prices was reduced by 55%, surpassing the cost reduction target.

Section 6.3: Deployment Strategy

Given the nature of the rental market, our model must be seamlessly integrable with existing property management systems and web services. We offer two primary deployment options:

1. **Docker Container:** A self-contained Docker image that can be deployed on the client's infrastructure, accepting POST requests for real-time predictions.
2. **REST API:** A locally deployable Flask-based REST API that allows easy integration with the client's existing applications, handling POST requests for predictions.

To facilitate user interaction and simplify integration, we also provide a basic Gradio application interface. This tool allows users to input data and receive predictions through the Flask API, offering a straightforward way to interact with the model.

Deployment Steps:

1. **Containerization:** Using Docker, the model and all dependencies are packaged into a container.
2. **API Setup:** The Flask API is deployed on a server, ready to handle incoming prediction requests.
3. **Integration:** The Docker container or Flask API is integrated into the client's system, with minimal disruption to existing operations.
4. **User Interface:** A Gradio application is set up to provide an easy-to-use interface for end users to interact with the model.

Conclusion

The deployment of the rental rate prediction model is designed to be efficient and easily integrable, ensuring minimal disruption to existing operations while providing accurate and timely predictions. By leveraging lightweight hardware requirements and flexible deployment options, the model is well-suited to meet the practical demands of Equity Residential and enhance their operational efficiency. Continuous monitoring and iterative improvements will further refine the model's performance and economic impact.