






# Report 3

 Assigned	 Ninel Yunusova
 Projects	 <u>Phase III - Model engineering</u>
 Status	Done

## Chapter 4: Model engineering

### Section 4.1. Literature research on similar problems

#### Paper 1: Predicting Real Estate Prices Using Machine Learning

**Authors:** K.A. Monson and L.J. Benjamin

**Publication:** Journal of Real Estate Research, 2020

#### Summary:

Monson and Benjamin applied various ML algorithms, including linear regression, decision trees, and gradient boosting machines, to predict real estate prices in urban areas. Their dataset included features such as square footage, number of bedrooms and bathrooms, location (zip code), and year built. The study found that gradient boosting machines outperformed other models with a root mean squared error (RMSE) of 24,000 compared to 32,000 for linear regression. The authors emphasized the importance of feature engineering and data preprocessing in improving model performance.

#### Key Insights:

- Gradient boosting machines provided superior performance.
- Feature engineering and preprocessing significantly impact the model's accuracy.
- The use of location-based features (zip codes) proved to be highly influential in predicting prices.

#### Study 2: Machine Learning Approaches for Housing Price Prediction

**Authors:** T. Zhang, Y. Liu, and W. Chen

**Publication:** International Journal of Computer Applications, 2019

## **Summary:**

Zhang et al. explored several ML models, including support vector machines (SVM), neural networks, and ensemble methods, to forecast housing prices. The dataset included diverse features such as property type, condition, neighborhood amenities, and historical price trends. Their results indicated that ensemble methods, particularly random forests, yielded the best performance with an RMSE of 20,000. The study also highlighted the utility of incorporating historical price trends and neighborhood amenities as predictive features.

## **Key Insights:**

- Ensemble methods, particularly random forests, showed the best performance.
- Historical price trends and neighborhood amenities are critical features for accurate predictions.
- The model's success heavily relied on the diversity and quality of the input features.

The insights from these studies will be instrumental in guiding the subsequent phases of our project. Key takeaways include the following:

### **1. Model Selection:**

- Considering the superior performance of ensemble methods (e.g., gradient boosting, random forests) and deep learning models in similar studies, these techniques will be prioritized in our model selection process.

### **2. Feature Engineering:**

- The importance of diverse and quality features, such as location-based features, historical price trends, and neighborhood amenities, will be emphasized. We will ensure comprehensive feature engineering to capture the complexity of the housing market.

### **3. Data Preprocessing:**

- Effective preprocessing techniques, as highlighted in the literature, will be applied to enhance the model's predictive power. This includes handling missing values, scaling features, and encoding categorical variables.

### **4. Baseline Performance:**

- The performance metrics from these studies (e.g., RMSE values) will serve as benchmarks for our model evaluating, providing a reference point to assess the efficacy of our models.

## Section 4.2. Define quality measures of the model

### Performance Metrics

#### 1. Root Mean Squared Error (RMSE):

- **Definition:** RMSE measures the average magnitude of the errors between predicted and actual values.
- **Reason for Use:** RMSE provides a clear metric of prediction accuracy, reflecting the standard deviation of the prediction errors. It is a commonly used metric for regression tasks and offers intuitive understanding of model performance.

#### 2. Mean Absolute Error (MAE):

- **Definition:** MAE calculates the average absolute differences between predicted and actual values.
- **Reason for Use:** MAE is less sensitive to outliers compared to RMSE, providing a more robust measure of prediction accuracy, especially when dealing with skewed data distributions.

We will use MSE for training, but to determine the best models and validation, we will use MAE to understand how many dollars our model is wrong.

## Section 4.3. Model Selection

### Input and Output Dimensions:

- **Input:** The input features include square footage, number of bedrooms and bathrooms, location (zip code), year built, property type, condition, neighborhood amenities, historical price trends, and economic indicators.
- **Output:** The output is a single continuous value representing the predicted apartment price.

#### 1. Decision Tree Regressor

##### Description:

The Decision Tree Regressor is a non-parametric model that splits the data into subsets based on feature values, creating a tree-like structure. Each node

represents a decision rule, and each leaf represents a predicted value. This model is intuitive and provides good explainability.

## **2. Gradient Boosting Regression**

### **Description:**

Gradient Boosting Regression builds an ensemble of trees sequentially, where each tree attempts to correct the errors of the previous ones. This model is powerful for regression tasks and often achieves high accuracy. However, it is less interpretable than a single decision tree.

## **3. First Neural Network (NN)**

### **Description:**

A Simple Neural Network consists of input layers, hidden layers, and an output layer. This model captures non-linear relationships between features and target variables. It is more flexible compared to Decision Trees but may require more data and tuning to perform well. Model have 3 hidden layers with different number of nervous in layer and ReLU function activation

## **4. Second Neural Network (NN)**

### **Description:**

A Simple Neural Network consists of input layers, hidden layers, and an output layer. This model captures non-linear relationships between features and target variables. It is more flexible compared to Decision Trees but may require more data and tuning to perform well. Model have 3 hidden layers with different number of nervous in layer and LeakyReLU function activation and dropout layers ( $p=0.5$ )

## **Section 4.4. Incorporate domain knowledge**

The chosen quality metrics—Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE)—are highly relevant to the task of predicting apartment prices. Both metrics focus on the accuracy of the predictions, which is crucial for providing reliable price estimates in the real estate market.

### **1. Root Mean Squared Error (RMSE):**

- **Definition:** RMSE measures the square root of the average squared differences between predicted and actual values.
- **Relevance:** RMSE is sensitive to larger errors, making it a useful metric when significant deviations from actual prices are particularly undesirable. It provides a clear indicator of the model's prediction

accuracy, directly impacting business decisions based on price forecasts.

## 2. Mean Absolute Error (MAE):

- **Definition:** MAE calculates the average absolute differences between predicted and actual values.
- **Relevance:** MAE is a straightforward measure of prediction accuracy, less affected by outliers than RMSE. It gives an intuitive understanding of average prediction errors, which is important for practical applications where consistent accuracy is valued.

**RMSE and MAE** are chosen for their relevance to the business objective of accurate price prediction. They directly measure the accuracy of the model's predictions, ensuring that the model meets the practical requirements of the real estate market.

## Section 4.5. Model training

### Train-Test Split Strategy

We employed an 80/20 split for training and testing datasets. This approach ensures that the model is trained on a substantial portion of the data while reserving a sufficient amount for evaluation, providing a realistic measure of its performance on unseen data.

- **Training Set:** 80% of the dataset
- **Testing Set:** 20% of the dataset

Additionally, we used k-fold cross-validation with  $k=4$ . This technique involves partitioning the training data into four subsets, training the model on three subsets, and validating it on the remaining subset. This process is repeated four times, with each subset used once as the validation set. The results are averaged to provide a robust estimate of the model's performance.

### Modeling Results and Comparisons

#### Decision Tree Regressor

We experimented with several hyperparameters for the Decision Tree Regressor to identify the best configuration.

- **Hyperparameters:**
  - **max\_depth:** [5, 10, 20]

- **max\_features:** [0.9, "sqrt", "log2"]
- **criterion:** ["friedman\_mse", "absolute\_error", "poisson"]
- **Results:**
  - **Best MAE on Training:** ~10
  - **Best MAE on Evaluation:** 10.5

The Decision Tree Regressor showed robust performance with relatively low MAE values, indicating its suitability as a baseline model.

### **Gradient Boosting Regression**

Gradient Boosting Regression was tested with a range of hyperparameters to enhance performance.

- **Hyperparameters:**
  - **loss:** ["squared\_error", "absolute\_error", "huber"]
  - **learning\_rate:** [0.01, 0.1, 0.2]
  - **subsample:** [0.5, 0.8, 1.0]
- **Results:**
  - **Best MAE on Training:** ~150
  - **Best MAE on Evaluation:** 156.4

While Gradient Boosting Regression had higher MAE values compared to the Decision Tree Regressor, it demonstrated good potential for improvement with further tuning.

### **Neural Network Models**

Two neural network configurations were tested to explore non-linear relationships in the data.

#### **First Neural Network:**

- **Architecture:** 3 fully-connected layers with ReLU activation
- **Results:**
  - **Best MAE:** ~3000

#### **Second Neural Network:**

- **Architecture:** 3 fully-connected layers with leaky ReLU activation and dropout regularization
- **Results:**
  - **Best MAE:** ~2400

Both neural network models underperformed compared to the tree-based models. The second neural network showed some improvement over the first, indicating that regularization techniques like dropout can enhance performance. However, further experimentation with hyperparameters and network architecture is necessary.

## Discussion

The choice of quality metrics (RMSE and MAE) was crucial in evaluating model performance, directly aligning with the business objective of accurate apartment price predictions. The Decision Tree Regressor emerged as the most effective model, achieving the lowest MAE values on both training and evaluation datasets. The Gradient Boosting Regression model, while not outperforming the Decision Tree Regressor, showed promise with moderate MAE values and potential for improvement through further tuning.

Neural networks, despite their capability to model complex non-linear relationships, did not perform as well in this task. This suggests that simpler, more interpretable models like decision trees might be better suited for this specific application, or that the neural networks require more extensive hyperparameter optimization and architecture adjustments.

## Section 4.6 Assure reproducibility

Reproducibility is a critical aspect of machine learning projects, ensuring that models can be consistently recreated and validated by others. This section covers the reproducibility of the Decision Tree Regressor model, focusing on method reproducibility, result reproducibility, and experimental documentation.

### Result Reproducibility

To validate the mean performance and assess the variance of the model using different random seeds. This practice ensures the robustness of the model and highlights any sensitivity to changes in the dataset split or initialization.

### Experimental Results:

- **Mean Absolute Error (MAE) on 6 samples of seed:**
  - Seed 1: 13.98

- Seed 2: 14.73
- Seed 3: 13.54
- Seed 4: 14.35
- Seed 5: 15.39
- Seed 6: 17.64

**Statistical Analysis:**

- **Average MAE:** 14.94
- **Variance of MAE:** 1.79

The average MAE and its variance across different seeds provide a clear indication of the model's performance and its consistency.