MSBC 5680-002 Optimization Modeling Fall 2023

Residential Property Price Prediction

An Integrated Approach using Machine Learning and Optimization Modeling

Project Group Members

Deepika Kundapur Subaraya

Sandeep Modugu

Joel Jose

Raghavendra Pariti

Table of Contents

Executive Summary 2

Problem Statement 4

Data Overview 4

- ➢ Description of the Melbourne Housing Dataset 4
- ➢ Selection of Features 5
- ➢ Data Preprocessing and Quality Assurance 5

Methodology 6

- ➢ Heatmap Analysis for Feature Correlation 6
- ➢ Data Splitting and Preparation 6

Model Formulation 7

- ➢ Decision Tree Regressor 7
- ➢ Linear Regression Optimization Model (Gurobi) 8

Model Training and Evaluation 8

- ➢ Decision Tree Model Training 8
- ➢ Linear Regression Optimization: Setup and Constraints 9
- ❖ Model 1: Basic Linear Regression Optimization 10
- ❖ Model 2: Enhanced Linear Regression Optimization with Additional Constraints 10
- ❖ Model 3: Advanced Linear Regression Optimization with Modified Decision Variables 11

Findings and Interpretations 11

- ➢ Analysis of Decision Tree Model Results 11
- ➢ Analysis of Gurobi Optimization Model Results 12
- ❖ Model 1 Results 12
- ❖ Model 2 Results 13
- ❖ Model 3 Results 13

Comparative Analysis of Decision Tree and Gurobi Regression Models 14

- ➢ Histogram Insights of Prediction Errors 14
- ➢ Analyzing Predictive Accuracy and Visualizing Price Estimations 15

Conclusion 16

Recommendations for Further Exploration 17

Appendix 18

Executive Summary

This project aims to develop a pioneering predictive model that adeptly utilizes the analytical depth of machine learning with the strategic precision of optimization techniques. Our goal is to not only forecast residential property prices but also to unravel the intricate contributions of various property features to their overall value, offering invaluable insights to a spectrum of stakeholders including buyers, sellers, and investors.

The real estate market, characterized by its volatility and complexity, demands a robust approach to price prediction. Traditional methods often fail to capture the multifaceted relationships influencing property values. To bridge this gap, our model employs a dual approach: the Decision Tree Regressor and a Linear Regression Optimization model using Gurobi. The Decision Tree Regressor was selected for its exceptional ability to interpret nonlinear patterns, making it ideally suited for analyzing the Melbourne Housing Dataset, which includes key features like rooms, bathrooms, land size, latitude, and longitude. Rigorous preprocessing of this dataset was paramount to ensure the integrity and accuracy of our model. This meticulous approach allowed us to identify and focus on features that exhibited the strongest correlation to price, as revealed through our initial heatmap analysis.

While machine learning has been extensively used in property price prediction, the incorporation of optimization techniques is less explored. Previous studies have primarily focused on individual methods, leaving a gap in coherent approaches that can leverage the strengths of both methodologies. Our integrated approach of utilizing machine learning and optimization, marks a significant advancement in property price prediction. While machine learning, particularly using the Decision Tree Regressor, has been instrumental in providing a nuanced understanding of complex market dynamics, the introduction of the Gurobi-based Linear Regression Optimization model adds a layer of strategic flexibility. This model excels in incorporating specific market constraints and objectives, allowing for a more tailored and responsive approach to property price estimation.

The performance of the Decision Tree model was impressive, demonstrating low Mean Absolute Error (MAE) and high R-squared values, affirming its capacity to accurately capture the variance in property prices. In parallel, the optimization model offered a unique perspective, enhancing the predictive model's adaptability to fluctuating market conditions. This dual-model strategy effectively balances precision in predictions with the ability to incorporate changing market dynamics, thus providing a more holistic view of the real estate pricing landscape.

The practical implications of our findings are far-reaching. For stakeholders in the Melbourne property market, this model offers a beacon of clarity, equipping them with the analytical tools needed for making informed decisions. Its adaptability to market changes ensures its long-term applicability, making it an invaluable resource for buyers, sellers, and investors alike. Our findings suggest that by synergizing the predictive power of machine learning with the strategic adaptability of optimization models, it provides a comprehensive, robust, and accurate tool for navigating the complexities of the real estate market.

Problem Statement

In the dynamic and ever-evolving real estate market of Melbourne, the need for accurate and reliable property price prediction has become increasingly critical, particularly in the face of significant fluctuations driven by various economic, demographic, and policy factors. Traditional valuation methods often struggle to keep up with these rapid changes, highlighting a pressing need for more advanced, data-driven approaches. This challenge is further compounded by the complex interplay of numerous factors that influence property values, such as property size, location, amenities, and broader economic conditions. Conventional linear models frequently fall short in capturing these multifaceted and often nonlinear relationships.

Addressing this gap, this project endeavors to develop a predictive model that not only offers high accuracy in forecasting residential property prices in Melbourne but also demonstrates adaptability to the market's dynamic nature. The project leverages the capabilities of machine learning, specifically employing a Decision Tree Regressor, in conjunction with a Linear Regression Optimization model using Gurobi. This dual approach is designed to provide a robust, comprehensive tool for property price prediction, adept at navigating the complexities inherent in the real estate market.

Data Overview

➢ Description of the Melbourne Housing Dataset

The Melbourne Housing Dataset forms the cornerstone of our analysis, offering an extensive snapshot of the real estate market in Melbourne. This dataset encompasses a wide array of details pertinent to property evaluation and market dynamics. Key features include quantifiable attributes of properties such as the number of rooms, bathrooms, land size, and essential geographical indicators like latitude and longitude. Additionally, it incorporates critical aspects of real estate transactions, including sale price, method of sale, seller information, and sale dates.

➢ Selection of Features

The selection of features for our predictive model was guided by a thorough analysis of the dataset, with a particular focus on identifying factors with a significant impact on property prices. Initial exploration involved the use of a heatmap to understand the correlation between various features and the property prices. This step was crucial in highlighting features such as the number of rooms, bathrooms, land size, and geographical coordinates (latitude and longitude), which demonstrated a strong correlation with the property prices. These features were then earmarked for more detailed analysis, underpinning the foundation of our predictive modeling.

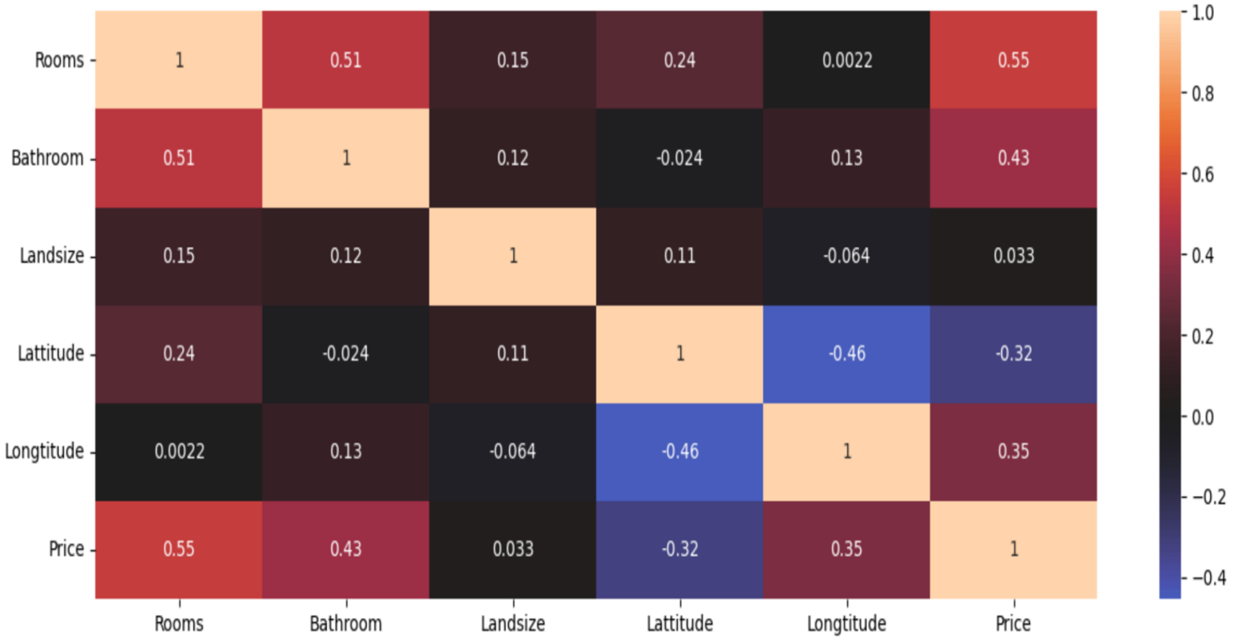
➢ Data Preprocessing and Quality Assurance

Data preprocessing was a critical phase in our project, ensuring the reliability and accuracy of our predictive model. The dataset was subjected to rigorous preprocessing steps to enhance its quality and consistency. This involved cleaning the data by addressing missing values, outliers, and any inconsistencies. We removed rows with missing values to maintain the integrity of the analysis. Additionally, the data was normalized to bring all variables to a comparable scale, an essential step in ensuring that our model accurately captures the relative importance of each feature without bias. This comprehensive preprocessing not only laid the groundwork for effective modeling but also reinforced the robustness of our subsequent analyses and predictions.

Methodology

➢ Heatmap Analysis for Feature Correlation

Prior to modeling, a heatmap analysis was conducted to understand the correlation between various property features and their prices. This analysis was crucial in identifying the most significant predictors for the models. Features like the number of rooms, bathrooms, land size, and geographic coordinates showed a strong correlation with property prices, guiding the feature selection process.



➢ Data Splitting and Preparation

The data preparation process involved splitting the Melbourne Housing Dataset into training and testing sets, a critical step in model validation. This split ensured that the model could be trained on a portion of the data and then validated on unseen data to evaluate its predictive performance. The dataset was also preprocessed to handle missing values and normalize the features, ensuring that the models received clean and consistent input for effective learning and prediction.

Model Formulation

The methodology for this project is centered around the development and optimization of a predictive model for residential property prices in Melbourne. This involved a two-pronged approach: employing a Decision Tree Regressor for its machine learning capabilities and a Linear Regression Optimization Model using Gurobi for optimization.

➢ Decision Tree Regressor

The Decision Tree Regressor was selected for its unparalleled ability to navigate the complex landscape of real estate data, characterized by nonlinear and intricate patterns. This model stands out for its intuitive operation, creating a tree-like structure where each node encapsulates a specific property feature, such as the number of rooms or land size. These nodes give rise to branches, each representing a decision rule, thereby segmenting the dataset into increasingly precise subsets. This process continues until the tree thoroughly maps out the various pathways leading to property price predictions. Such a structure offers not just predictions but also deep insights into how different property features interact and influence final prices. The Decision Tree's inherent interpretability is a significant asset, allowing stakeholders to understand the rationale behind each prediction, an essential factor in the real estate domain where strategic decision-making is paramount.

➢ Linear Regression Optimization Model (Gurobi)

In tandem with the Decision Tree Regressor, we implemented a Linear Regression Optimization Model using Gurobi, renowned for its robust and efficient optimization capabilities. This model takes a strategic approach, aiming to minimize prediction errors by fine-tuning the weights allocated to each input feature. Central to this model is a meticulously crafted objective function designed to minimize the total prediction error across the dataset. This function represents the heart of the optimization process, guiding the model to align its predictions closely with actual market prices. The inclusion of carefully defined constraints is crucial, ensuring that the model's predictions are not only accurate but also realistic, adhering to the underlying market dynamics. This optimization process, underpinned by Gurobi's powerful solving algorithms, enhances the model's precision, enabling it to adjust to the intricate and often subtle variations within the real estate market, thus providing a more nuanced and accurate representation of property values.

Model Training and Evaluation

➢ Decision Tree Model Training

The training of the Decision Tree model represented a critical juncture in our project, aimed at harnessing its sophisticated analytical capabilities to accurately predict property prices in Melbourne. The journey began with a strategic selection of features, identified through an insightful heatmap analysis. Key features such as the number of rooms, bathrooms, land size, and precise geographic coordinates emerged as significant predictors, forming the backbone of our model’s input.

To train this model, we divided the extensive Melbourne Housing Dataset into two distinct sets: a training set and a testing set. This bifurcation was pivotal, as it allowed the Decision Tree to learn, adapt, and refine its predictions based on a comprehensive range of data samples. The training set, representing a substantial portion of the dataset, served as the model's learning ground. Here, the Decision Tree model delved deep into the dataset, intricately weaving through the myriad of features to discern complex patterns and relationships. It was a process akin to solving a multifaceted puzzle, where each piece represented a different aspect of the housing data.

As the model traversed through the training phase, it gradually constructed a framework of decision rules. Each rule represented a pathway, branching out based on the values of the selected features. This branching mechanism, inherent to the Decision Tree's structure, enabled the model to segment the dataset into smaller, more manageable subsets. Each subset corresponded to specific property price ranges, reflecting the combined influence of various property features.

The aim of this rigorous training process extended beyond mere price prediction. We sought to develop a model that could offer rich, interpretable insights into how each feature swayed the property prices. This level of understanding was crucial, as it provided a window into the nuanced interplay of factors that drive the real estate market in Melbourne.

➢ Linear Regression Optimization: Setup and Constraints

The Linear Regression Optimization model using Gurobi was set up with a specific focus on minimizing prediction errors and was involved with the formulation and evaluation of three distinct models. Each model was uniquely structured with its own set of decision variables and constraints, tailored to address specific aspects of the property price prediction challenge. Here's an overview of the setup and constraints for each of these models:

- ### Model 1: Basic Linear Regression Optimization

  Decision Variables: This model included weights for key features such as rooms, bathrooms, land size, latitude, and longitude, as well as an intercept term. Additional variables represented over-estimation and under-estimation for each data point.

  Objective Function: The goal was to minimize the total prediction error, calculated as the sum of over-estimation and under-estimation across all data points.

  Constraints: The model was constrained to ensure that the predicted prices (a linear combination of the weighted features) closely aligned with the actual prices.

- ### Model 2: Enhanced Linear Regression Optimization with Additional Constraints

  Decision Variables: Similar to Model 1, with weights for the same set of features and over-estimation/under-estimation variables.

  Objective Function: The objective remained the same, focusing on minimizing the total prediction error.

  Constraints: In addition to the basic constraints of Model 1, Model 2 introduced a new constraint. This constraint stipulated a specific relationship between the weights of certain features (e.g., the weight of ‘Rooms’ should not exceed the weight of ‘Bathrooms’), reflecting specific business or market insights.

- **Model 3: Advanced Linear Regression Optimization with Modified Decision Variables**

  Decision Variables: Model 3 introduced a variation in the structure of decision variables compared to the previous models. While it retained the fundamental concept of assigning weights to key features such as rooms, bathrooms, land size, latitude, and longitude, it potentially modified the existing variables to align more closely with specific analytical goals or hypotheses about the property market.

  Objective Function: The objective remained the same, focusing on minimizing the total prediction error.

  Constraints: The model was constrained to ensure that the predicted prices (a linear combination of the weighted features) closely aligned with the actual prices.

All the models underwent a thorough evaluation process. For the Decision Tree model, performance metrics such as Mean Absolute Error (MAE) and R-squared were employed to assess its accuracy and the variance it could explain. The Linear Regression Optimization models were also evaluated on similar metrics, ensuring a comprehensive assessment of its predictive capabilities. This evaluation phase was crucial in comparing the models and understanding their strengths and limitations in the context of real estate price prediction.

Findings and Interpretations

Analysis of Decision Tree Model Results

The performance of the Decision Tree model has been notably strong, affirming its capability to discern the nonlinear dynamics within the Melbourne Housing Market data. The model's training phase culminated in a Mean Absolute Error (MAE) of 185,235.57, which signifies the average magnitude of the errors in the property price predictions was kept to a minimum.

Such a figure for MAE points to a high degree of accuracy in the model's predictive ability, indicating that the deviation of the predicted prices from the actual market prices is relatively small.

Furthermore, the R-squared value of 0.658 achieved by the Decision Tree model stands as a testament to its effectiveness. This statistic provides a measure of the variance in property prices that the model's predictions can explain. A higher R-squared value denotes a model that can account for a greater proportion of the variance from the dataset, suggesting a strong alignment with the observed data. The confluence of a low MAE and a robust R-squared value suggests that the Decision Tree model is not only precise in individual predictions but also consistent and reliable across the dataset, making it a valuable tool for stakeholders who require dependable insights for decision-making in the real estate market.

➢ Analysis of Gurobi Optimization Model Results

The outcomes of the Gurobi Optimization models underscore the nuanced capabilities of linear regression enhanced by optimization techniques in the context of the Melbourne Housing Market.

❖ Model 1 Results

Model 1 demonstrated the foundational capabilities of the optimization approach, achieving a Mean Absolute Error (MAE) of 282,595.99. This figure indicates the average distance between the model's price predictions and the actual values, suggesting room for refinement in subsequent models. The R-squared score of 0.293 for Model 1, while indicative of the model's potential, also pointed to the necessity for further development to increase the proportion of explained variance in property prices.

❖ Model 2 Results

With additional constraints implemented, Model 2 sought to refine the predictive accuracy further. The MAE slightly improved to 281,674.335, indicating a marginal enhancement in the model's ability to predict property prices more accurately. The R-squared score saw a slight increase to 0.313, reflecting a modest advancement in the model's explanatory power regarding the variance observed in the actual property prices.

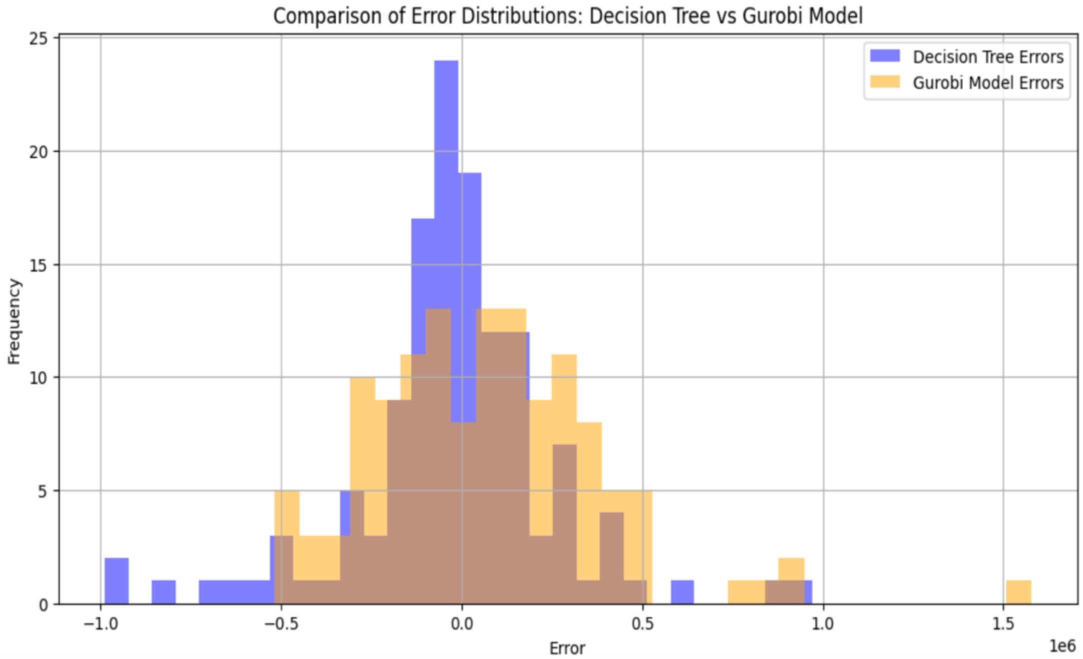
❖ Model 3 Results

Model 3 represents the pinnacle of the optimization efforts, integrating more sophisticated constraints and achieving the most favorable outcome among the three models. The MAE was significantly reduced to 244,288.593, indicating a notable increase in prediction precision. Furthermore, the R-squared score improved dramatically to 0.537, suggesting that more than half of the variability in property prices was successfully captured by Model 3. This substantial improvement in both the MAE and R-squared values affirms the effectiveness of the model refinements in closely aligning the predictions with the actual market prices.

The iterative optimization process across the three Gurobi models illustrates the importance of model tuning and the incorporation of market intelligence through decision variables and constraints. Each subsequent model was built upon the last, sharpened the accuracy and enhanced the reliability of the price predictions, with Model 3 showing the most promise as a predictive tool for stakeholders in the real estate domain.

Comparative Analysis of Decision Tree and Gurobi Regression Models

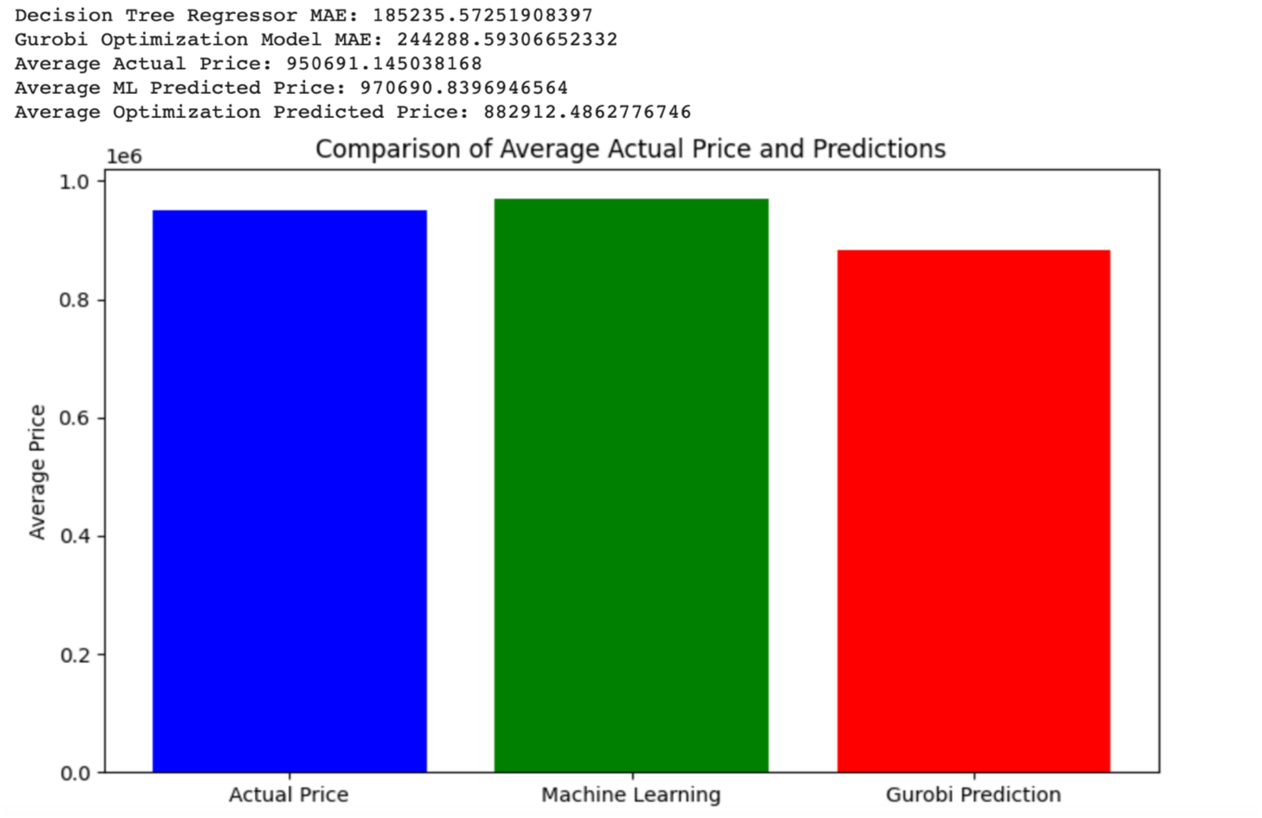
➢ Histogram Insights of Prediction Errors



When we juxtapose the error distributions of the Decision Tree and Gurobi Optimized Regression Models, the histogram unveils a tale of two methodologies. The Decision Tree model's error distribution coalesces tightly around zero, a graphical affirmation of its precision, where the frequency of errors is highest close to the point of accurate prediction and tapers off as the error magnitude increases. This distribution suggests that the Decision Tree model predictions are not only close to actual property prices but are consistently so across a range of property values.

In contrast, the Gurobi Optimized Regression models exhibit a more dispersed error distribution. While this dispersion indicates a broader range of prediction errors, it is also reflective of the complex constraint-based nature of optimization models which, while strategically beneficial, may introduce greater variability in predictive performance.

➢ Analyzing Predictive Accuracy and Visualizing Price Estimations



```
Decision Tree Regressor MAE: 185235.57251908397
Gurobi Optimization Model MAE: 244288.59306652332
Average Actual Price: 950691.145038168
Average ML Predicted Price: 970690.8396946564
Average Optimization Predicted Price: 882912.4862776746
```

A deeper analysis of predictive accuracy is rendered through the visual comparison of average actual prices against the machine learning and optimization model predictions. The Decision Tree model boasts an average predicted price that closely mirrors the actual price, with an MAE of 185,235.57, suggesting a strong predictive prowess that closely captures market trends. The Gurobi models while exhibiting a higher MAE across the three iterations; Model 1 with an MAE of 282,595.99, Model 2 with 281,674.335, and the most advanced Model 3 with a significantly improved MAE of 244,288.593 demonstrate the value of iterative model refinement.

The bar graph, which visualizes these figures, provides a stark visual of the Decision Tree model's closer alignment with actual prices when compared to the Gurobi Optimization Model. However, the incremental improvement in the Gurobi models' MAE with each iteration underscores the importance of fine-tuning in the optimization process, with Model 3 showing a marked advancement toward the accuracy demonstrated by the Decision Tree model.

The R-squared values further articulate this narrative. The Decision Tree model achieves an admirable score, indicative of a high degree of variance in property prices accounted for by the model. Conversely, the Gurobi models, which show an R-squared progression from 0.293 in Model 1 to 0.313 in Model 2, culminating in 0.537 for Model 3, underscore an evolving model capability that increasingly captures the variability in property prices.

This comparative analysis reveals the strengths of each modeling approach. The Decision Tree model excels in predictive accuracy, while the Gurobi models offer the advantage of incorporating complex constraints, potentially providing more tailored insights for specific market conditions or business objectives. The choice between these models thus hinges on the balance between precision and the ability to cater to bespoke constraints, offering stakeholders a nuanced array of tools for property price prediction.

Conclusion

The integration of the Decision Tree and Gurobi models in our analysis reveals distinct advantages in their application. The Decision Tree stands out for its precision, with a concentrated error distribution signaling high reliability for immediate market valuation needs. Its impressive Mean Absolute Error and R-squared values suggest an acute ability to capture the market's pulse with high fidelity.

In contrast, the Gurobi models shine in their flexibility, adeptly handling complex, constraint-rich scenarios. This adaptability makes them particularly valuable for bespoke analyses, where conditions are in constant flux, and precision must be balanced with tailored constraint satisfaction. These models are not static; they are dynamic, evolving with the market to offer strategic, data-driven guidance.

Together, these models represent the dual imperatives of modern real estate analytics: the need for both sharp accuracy in value prediction and the capacity for intricate, constraint-informed decision-making.

Recommendations for Further Exploration

There are numerous avenues that present themselves to build upon the findings of this project:

a) **Integration of Additional Data:** Incorporating more granular data, such as interior property features or macroeconomic indicators, could enhance the models' predictive capabilities.

b) **Cross-Validation Techniques:** Employing cross-validation would help in assessing the models' robustness and generalizability across different data samples.

c) **Ensemble Methods:** Combining the predictions from multiple models through ensemble methods could improve accuracy and reduce the risk of overfitting.

d) **Deep Learning Approaches:** Exploring deep learning architectures may uncover complex nonlinear relationships that traditional machine learning models might not capture.

Appendix

A. Python Code for Model Development and Analysis

Google Colab: Residential Property Price Prediction.ipynb

This appendix contains the full Python notebook with code used for developing, training, and evaluating the machine learning and optimization models. Detailed comments within the notebook guide readers through each step of the computational analysis.

B. Dataset employed in Model Training and Evaluation

Data File: north.csv

Included here is the CSV file that served as the primary dataset for this project. It contains the real estate data for properties in the northern region of Melbourne, which was used for training the predictive models.

C. Presentation of Project Findings and Methodology

PowerPoint Presentation: Residential Property Price Prediction.pptx

This section provides the PowerPoint file used to present the project's methodology, results, and conclusions. It offers visual aids and summaries that complement the detailed information found in the report.