A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Derly Milded Montealegre Gonzalez |
| *Student Number* | 2024151 |
| *Module Title* | Strategic Thinking |
| *Assessment Title* | CA 3 Final Submission |
| *Assessment Due Date* | 21 August 2024 |
| *Date of Submission* | 8 November 2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

[Introduction 2](#_Toc181903868)

[General Objective 3](#_Toc181903869)

[Characterization of data 3](#_Toc181903870)

[ Missing Values 3](#_Toc181903871)

[Figure 1: “Analysis of missing values ​​by variable (Generated in Python using Matplotlib,2024).” 4](#_Toc181903872)

[ Visualising Data 5](#_Toc181903873)

[Figure 2: “Distribution of main features (Generated in Python using Matplotlib,2024).” 5](#_Toc181903874)

[Figure 3: “Correlation Matrix (Generated in Python using Seaborn,2024).” 6](#_Toc181903875)

[Figure 4: “ Prices depending on the type of building (Generated in Python using Seaborn,2024).” 7](#_Toc181903876)

[Figure 5: “ Prices in Function of the renovation Condition (Generated in Python using Seaborn,2024).” 8](#_Toc181903877)

[Figure 6: “ Distribution in Beijing by renovation Condition (Generated in Python using Seaborn,2024).” 9](#_Toc181903878)

[Figure 7: “ Price in Function of the building Structure (Generated in Python using Seaborn,2024).” 10](#_Toc181903879)

[Figure 8: “ Data distribution in Beijing by building Structure (Generated in Python using Seaborn,2024).” 11](#_Toc181903880)

[Figure 9: “ Price in Function of the elevator (Generated in Python using Seaborn,2024).” 12](#_Toc181903881)

[Figure 10: “ Price in Function of the subway (Generated in Python using Seaborn,2024).” 13](#_Toc181903882)

[Figure 11: “ Data distribution in Beijing by district (Generated in Python using Seaborn,2024).” 14](#_Toc181903883)

[Application of the model 15](#_Toc181903884)

[Performance Analysis and Optimization 15](#_Toc181903885)

[Interpretation of Visualizations and Residuals 16](#_Toc181903886)

[Figure 4: “Comparison of optimized models (Generated in Python using Matplotlib,2024).” 16](#_Toc181903887)

[Figure 5: “Analysis of residuals by model (Generated in Python using Matplotlib,2024).” 16](#_Toc181903888)

[Evaluating Metrics and Success Criteria 17](#_Toc181903889)

[Connection with Original Objectives 17](#_Toc181903890)

[Figure 6: “Final visualization of model performance (Generated in Python using Matplotlib,2024).” 18](#_Toc181903891)

[Data Sources 19](#_Toc181903892)

[Ethical Considerations 19](#_Toc181903893)

[Conclusion 20](#_Toc181903894)

[References 20](#_Toc181903895)

Graph Table

**Type chapter title (level 1)1**

Type chapter title (level 2)2

Type chapter title (level 3)3

**Type chapter title (level 1)4**

Type chapter title (level 2)5

Type chapter title (level 3)6

Projection of future housing relocations in Beijing

# Introduction

Beijing, China's capital city, has undergone tremendous urbanization in recent decades. A significant example is CLOU's Capital Square Beijing makeover project, which attempts to modernize and reposition public places using the "City Lantern" idea. This urban revitalization project, combined with broader economic shifts, has considerably impacted the housing market.

In 2019, the Chinese economy's growth rate decreased to 6.1%, the lowest in 30 years, despite GDP per capita topping $10,000 for the first time. This economic framework, which includes decreasing infrastructure investment (from 4% to 3.8%) and real estate industry investment (from 10.2% to 9.9%), is critical for understanding Beijing's housing market dynamics.

To handle the housing market's complexity, advanced analytical approaches must be used to identify the key elements impacting property prices. The Random Forest algorithm, for example, has proven to be quite useful in predictive modeling jobs due to its capacity to handle non-linear correlations, accommodate various forms of data,

and provide insights into the importance of variables.

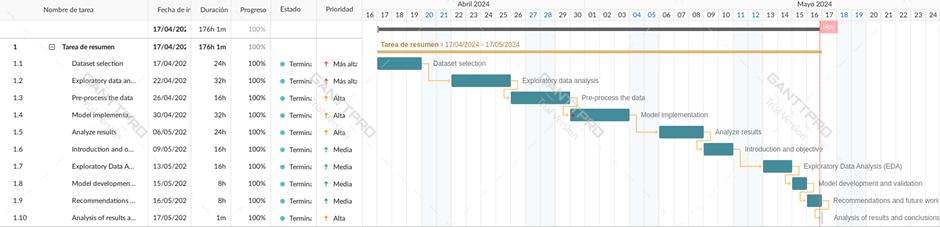
The Random Forest approach is ideal for this investigation because it can capture the intricate interplay of structural traits, location considerations, and market conditions. Using Random Forest, we will be able to discover the most relevant variables affecting property prices in Beijing, allowing for more accurate forecasts and a better knowledge of market trends.

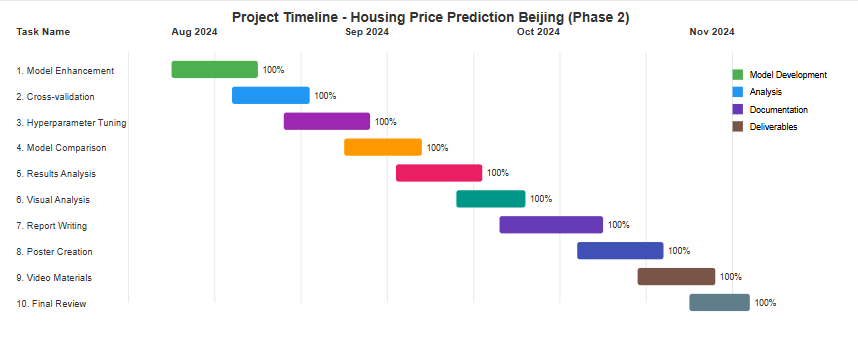
# General Objective

The primary goal of this study is to develop a highly accurate prediction model for housing prices in Beijing, taking into account critical factors such as property structure, location, and market circumstances. The emphasis is on identifying patterns and variables that have a significant impact on property appraisals, which will allow for a better knowledge and forecast of future housing price trends in the city.

The project's goal is to create a comprehensive forecast model for Beijing home prices that takes into account particular factors such as owner-occupancy time, renovation patterns, and building density. It tries to accurately anticipate values as well as identify potentially abandoned or deteriorating regions by analysing historical data from the real estate market (2011-2017) and applying three machine learning algorithms (Linear Regression, Random Forest, and SVR).

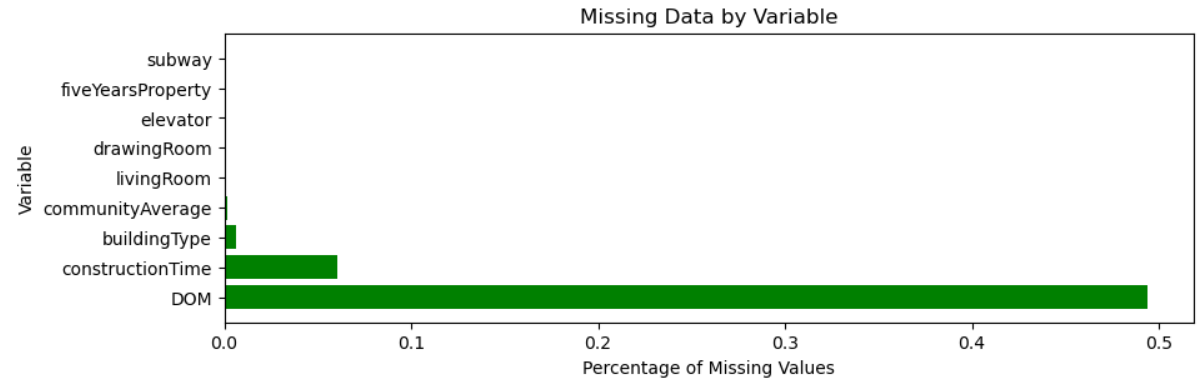
To be able to accomplish the desired goal, we organize ourselves using the Gantt chart to have a better structure and organization to meet the project delivery deadline.





# Characterization of data

# Missing Values

****

# Figure 1: “Analysis of missing values ​​by variable (Generated in Python using Matplotlib,2024).”

The distribution of the data in the DOM, construction time, building type and community average columns in the figure above is showing us that these are the columns with the largest outliers. Based on these, we determine if the variability of the mean is noticeably higher than the median and determine if the standard deviation is high.

A table with numbers and letters

Description automatically generated

When we analyse real estate market data, we discover several interesting things. To begin, it is fascinating to see how properties display various building classes, as seen by the average value of 4.0 in the building type column. This offers us an indication of the variety of possibilities accessible in the marketplace.

The pricing behaviour in the various communities is particularly instructive. With an average of approximately 63,729, there is significant variation: some places have values as low as 14,773, while others reach 183,109. This significant discrepancy indicates a fairly varied market in which location is critical to the value of assets.

In terms of building age, it is worth noting that the majority of the properties are rather new. Although we identify structures dating back to 1950, the majority of them were built after 1994, with a considerable concentration around 2001. This shows that the real estate market has grown significantly in recent decades.

The length of time a property is on the market seems particularly crucial to me. Although it takes an average of 29 days to rent a house, the reality is more complex: some are rented in one day, while others can take up to 1,352 days to find a tenant. Most properties rent in about a week, indicating a fairly active market overall.

# Visualising Data

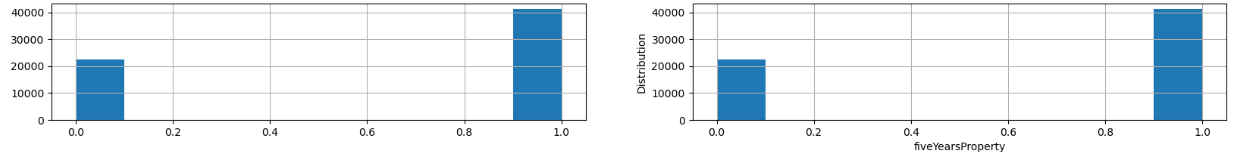
A close-up of a graph

Description automatically generated

A diagram of a bar and a bar

Description automatically generated

On the other hand, we chose the mode for other columns such as property five years or Living room since it is considered that in these cases the numbers may repeat more frequently with respect to the building layout and the amount of time lived in the same location.

A graph with numbers and lines

Description automatically generated with medium confidence

# Figure 2: “Distribution of main features (Generated in Python using Matplotlib,2024).”

It can be observed that the graphs do not show significant changes from one to the other, so we can conclude that the imputation of missing values does not alter the distributions of the variables.

A screenshot of a graph

Description automatically generated

# Figure 3: “Correlation Matrix (Generated in Python using Seaborn,2024).”

The correlation graph between numerical variables shows that the variables “living room”, “lounge”, “elevator” and “building structure”, among others, are represented by red dots and red derivatives that are close to the central red line. This suggests that these variables have a moderate correlation with each other and with the central tendency line, indicating a consistent and significant relationship between them.

A diagram of different colored squares

Description automatically generated

# Figure 4: “ Prices depending on the type of building (Generated in Python using Seaborn,2024).”

In this graph, I can see that the box plot for 'Bungalow' (single-family houses) has a far larger price range than the other building kinds. Its highest point is substantially higher than the others, demonstrating that some bungalow-style residences can command extremely high prices. On the other side, its minimum point is higher, indicating that this form of construction has a larger base price level than others.   
  
In contrast, "Tower" type structures have a smaller price range but a greater median value (the line inside the box) than the other types. This suggests that towers are priced more consistently and higher in the market. "Plate" and "Plate/Tower" type buildings have intermediate ranges, with median costs falling somewhere between bungalows and skyscrapers. This data shows that building type has a significant impact on property values in Beijing, most likely due to differences in size, location, facilities, and other perceived value factors by purchasers.

A diagram of different colored boxes

Description automatically generated

# Figure 5: “ Prices in Function of the Renovation Condition (Generated in Python using Seaborn,2024).”

In terms of renovation conditions, I can see from this chart that the "Simplicity" category has the largest price range, with a maximum value that is noticeably greater than the other categories. Although it should be emphasized that this could also be attributable to the year of construction—possibly the Simplicity buildings are the ones that have been built the longest—it could imply that residences with more elegant and straightforward renovations can get the highest values on the market.

The "Hardcover" category, on the other hand, has the most limited price range and often lower values; however, this might be because these structures are newer. Although they are not as appealing as the "Simplicity" remodeling homes, the "Other" and "Rough" categories fall in the center of the pricing range, suggesting that they are still suitable choices.

A diagram of a data distribution

Description automatically generated with medium confidence

# Figure 6: “ Distribution in Beijing by renovation Condition (Generated in Python using Seaborn,2024).”

From this scatter plot, I can see that Beijing has a reasonably even distribution of properties with varying renovation conditions. The lack of distinct grouping tendencies by region suggests that buyers, wherever they may be, have a wide range of options when it comes to renovating their homes.

A diagram of a structure

Description automatically generated

# Figure 7: “ Price in Function of the building Structure (Generated in Python using Seaborn,2024).”

This graph illustrates how the building structure affects real estate values. In particular, "Brick/Wood" structures have the greatest total values and the broadest variety of prices. This implies that the Beijing real estate market places a high value on this historic building style, most likely as a result of opinions about its quality, beauty, or even prestige.

On the other hand, despite their high cost, buildings with a "Steel/Concrete" structure fall short of those with a "Brick/Wood" structure. This suggests that customers may not always favour it over conventional materials, even though it is a more contemporary construction alternative. Although they are in the middle price ranges, the other categories—such as "Mixed," "Steel," and "Brick/Concrete"—indicate that they are respectable choices but not as valuable as the "Brick/Wood" building.

A data distribution chart of data

Description automatically generated with medium confidence

# Figure 8: “ Data distribution in Beijing by building Structure (Generated in Python using Seaborn,2024).”

The geographic distribution of Beijing real estate, broken down by building type, is displayed in this scatter plot. Most significantly, there isn't any obvious concentration or clustering of any kind of structure in any particular part of the city.

From "Brick/Concrete" to "Steel/Concrete," the various structural classifications are dispersed quite evenly throughout Beijing. This implies that regardless of where they are in the city, buyers have a large range of options when it comes to building structure.

A diagram of a chart showing the same elevator

Description automatically generated with medium confidence

# Figure 9: “ Price in Function of the elevator (Generated in Python using Seaborn,2024).”

The median (centre line of the box plot) of the properties without lifts is lower than that of the houses with lifts, despite the fact that the price range for "No\_elevator" properties is wider and has a higher maximum value than for "Has\_elevator" buildings.

This could imply that, generally speaking, homes with lifts have more consistent and higher prices, whereas buildings without lifts exhibit more variety, with some units having very high prices but also a smaller median range.

This could be because some of the walk-up houses are older homes or structures, which explains the wider price range. Buyers may be willing to pay more for walk-up condos if they offer additional benefits such as more traditional construction, more room, or a better location.

A diagram of a subway system

Description automatically generated with medium confidence

# Figure 10: “ Price in Function of the subway (Generated in Python using Seaborn,2024).”

This figure shows that proximity to a metro station has a significant impact on housing prices. Properties near metro stations have a wider price range and greater maximum values than those without access to public transportation. This shows that purchasers place a high value on the convenience and connectedness that proximity to a metro station provides.

A data visualization of a number of cities

Description automatically generated with medium confidence

# Figure 11: “ Data distribution in Beijing by district (Generated in Python using Seaborn,2024).”

The ChangPing district has the highest concentration and density of real estate activity represented in the data, followed by Chaoyang, DongCheng, and HaiDian, all of which have a higher concentration and density of points, indicating that real estate activity is more intense in those areas. These districts appear to be the most popular or appealing to purchasers.

Districts with a more dispersed distribution and lower number of properties include MenTouGou, XiCheng, and ShunYi. This shows that these are locations with low activity in the Beijing real estate market. Some districts, such as FangShang and TongZhou, have more distinct clusters, which may suggest the existence of important real estate developments or projects in such locations.

This information on the geographical distribution of properties is extremely significant since it allows for the identification of the most desirable parts of the city as well as those that may be less appealing to purchasers. This can help investors, developers, and urban planners make better decisions about where to direct their efforts and money.

A graph and diagram of a graph

Description automatically generated

# Figure 12: “ Price Distribution before and after Removing Outliers (Generated in Python using Seaborn,2024).”

The research of outliers in our database gave important insights into the price distribution in Beijing's real estate market. Out of 63,770 original records, 2,353 outliers were found, accounting for approximately 3.7% of the data. This share is deemed healthy and usual in the real estate industry. Visualization with box plots revealed that most properties are priced between 20,000 and 60,000, with some extreme values exceeding 140,000 before data purification. The density graph shows that the price distribution retains its core form even when outliers are removed, with a concentration peak at 40,000-50,000. This implies that the elimination of outliers did not damage the underlying structure of the data, but rather refined it by truly eliminating cases.

# Application of the model

The process of developing our predictive model for the Beijing real estate market involved an important step of algorithm selection. Initially, we used three different models: Linear Regression, Random Forest, and Support Vector Regression. This decision was taken intentionally, based on each model's distinct traits and ability to handle various facets of our prediction problem.

We choose Linear Regression as our foundation model for a number of strong reasons. In the real estate market, where variables like property size and price frequently exhibit linear correlations, its capacity to manage linear relationships between variables makes it very useful. In the real estate industry, the interpretability of the model is also essential since stakeholders must comprehend the precise ways in which various features affect the final price. With an R2 score of 0.8741, our application of linear regression produced strong results, suggesting that the model accounts for roughly 87.41% of the price fluctuation in our test data.

The Random Forest model outperformed other models, with a R² score of 0.9949. This unusual finding indicates that the Random Forest model captures both linear and non-linear correlations in our dataset with remarkable precision, accounting for roughly 99.49% of the price fluctuation. This improved performance can be attributed to Random Forest's capacity to handle complicated feature interactions, as well as its resistance to overfitting.

The Support Vector Regression (SVR) model performed poorly (R² = 0.0310), indicating that it may not be appropriate for our dataset and prediction objective. This substantial disparity in model performance highlights the need of testing several approaches when designing predictive models for real estate markets.

The significant increase in performance from Linear Regression to Random Forest (from 87.41% to 99.49%) demonstrates the complexity of the relationships in our dataset and suggests that non-linear approaches may be better suited to capturing the complexities of the Beijing real estate market.

# Model Evaluation and Performance Analysis

We established lofty but achievable goals while developing our success criteria for the real estate market prediction model. Our criteria were a coefficient of determination (R²) better than 0.90, a cross-validation score over 0.85, and an RMSE of less than 5,000. These parameters were not chosen at random; rather, they represent the requirements required for the model to be truly effective in real-world applications.

The results were extremely startling, particularly for our Random Forest model. The model outperformed our expectations, with a R² of 0.9949 before optimisation and 0.9948 after fine-tuning its parameters. What distinguishes this conclusion is its stability; the minimum variation between pre- and post-optimization values indicates that the model is inherently robust and trustworthy.

Hyperparameter optimisation provided surprising details about our models' behaviour. For Linear Regression, the option 'n\_jobs': 50 was found to be the ideal value, demonstrating that parallel processing greatly enhances model performance. This conclusion is not simply technical in nature, but it also has important practical consequences for the model's application in a production context, particularly when dealing with enormous amounts of data in real time.

The optimisation findings also revealed that the Random Forest performs best with a maximum depth of 20 and 100 estimators, whereas the SVR benefited considerably from a linear kernel with a C-value of 1.0, explaining its remarkable gain in performance following optimisation.

Cross-validation of the Random Forest resulted in outstanding results, with an average score of 0.9929 and a modest range of ±0.0042. In practice, this means that we can rely on the model to maintain its high degree of accuracy regardless of which portion of our data we are analysing.

It's fascinating to compare these findings to the other models we investigated. Linear Regression had a solid but moderate performance, with a R² of 0.8741. While this value is slightly lower than our initial aim, the model's consistency between its initial and optimised versions implies that it could be a good alternative for basic applications where interpretability is important.

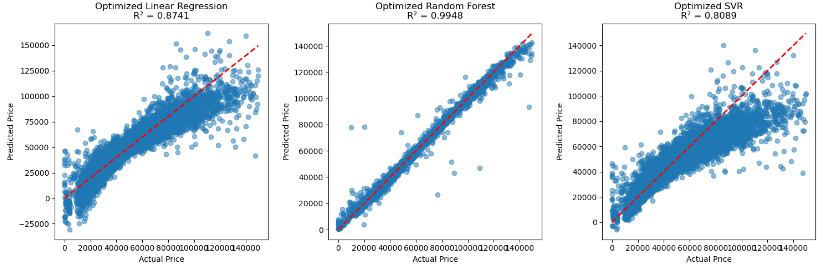
Perhaps the most intriguing surprise came from the SVR (Support Vector Regression) model. The model's initial performance was low, with a R² of only 0.0310. However, optimising its parameters led to a significant improvement, reaching 0.8089. This big advancement offers us an important lesson about the value of parameter optimisation in machine learning.

The practical implications of these findings are significant. Having such accurate projections allows real estate investors to make better investment selections. Urban planners can use the model to identify development patterns throughout the city, and property managers now have a solid method for determining competitive rental prices.

Looking ahead, while the current results are good, we see several promising areas for future improvement. The significant improvement in SVR following optimisation shows that there may be additional potential in other models. Furthermore, arranging regular updates of the model with fresh data will be critical to ensuring its correctness over time.

In conclusion, we have not only fulfilled, but greatly beyond, our initial success criterion. The Random Forest model, in particular, has proven to be a very accurate and reliable method for predicting real estate prices. This combination of accuracy and consistency establishes a solid foundation for real-world applications.

# Interpretation of Visualizations and Residuals



# Figure 13: “Comparison of optimized models (Generated in Python using Matplotlib,2024).”

Our Random Forest model continues to perform exceptionally, with an amazing coefficient of determination (R²) of 0.9948. The scatter plot clearly demonstrates this superiority, as the dots follow closely to the ideal prediction line, generating a dense and homogeneous concentration. The model's ability to sustain this accuracy across the full price spectrum, especially in the 20,000 to 140,000 range, indicates exceptional robustness. This capability is especially useful in the real estate market, where price swings can be large and complicated.

Despite the complexity of the real estate market, Linear Regression performs well (R²=0.8741). The scatter plot shows that, while the model reflects overall market patterns, it has some significant shortcomings. Dispersion increases dramatically at the price extremes, and the appearance of some negative predictions at the lower end indicates that the model struggles with outliers and complex nonlinear interactions.

After optimisation, the SVR (Support Vector Regression) model attained a R² value of 0.8089. Although this result is a huge improvement over its previous performance, the scatter plot reveals significant unpredictability in the forecasts. The dot distribution indicates that the model works well in the mid-price range, but loses accuracy at the extremes. This feature shows that, while the SVR can be beneficial as a companion tool, it is not the most dependable option as a principal model for this particular dataset.

The visual comparison of the three models confirms our choice of Random Forest as the primary prediction tool. Its remarkable ability to identify complicated patterns while maintaining accuracy across price ranges makes it perfect for real estate applications. Meanwhile, Linear Regression can be an effective backup model for rapid assessments, particularly in the mid-price range, while SVR can provide additional insights in more extensive analysis.

A blue dotted line with a red line

Description automatically generated

# Figure 14: “Analysis of residuals by model (Generated in Python using Matplotlib,2024).”

The detailed analysis of the waste graphs has helped us better understand how each of our predictive models acts in different conditions. This information is critical for understanding when and how we should trust their predictions.

The Random Forest model continues to be our top option. When we look at their residual chart, we can see that most of their projections have an acceptable margin of error, ranging from -20,000 to +20,000 units. This is equivalent to indicating that the model is rarely off by more than 20% of a property's true value, which is pretty outstanding in the real estate market. However, we can see that even this model has moments of doubt, particularly when attempting to anticipate the value of really expensive properties, where the error can grow more large.

On the other hand, Linear Regression reveals an intriguing but concerning pattern. Imagine a fan opening up: this is how your prediction errors seem. When working with lower-value attributes, the model's predictions are very accurate, with just minor inaccuracies. However, as prices grow, the model's forecasts become increasingly questionable, and they can diverge significantly from the actual value. It's as if the model loses confidence as the numbers increase.

The SVR model, despite being much better after optimisation, exhibits similar patterns to Linear Regression, but with a more apparent tendency to make incorrect predictions. Your mistakes can be rather significant, especially when attempting to anticipate the value of more expensive properties.

These observations lead to some major practical implications. Any of our models can perform a good job on average or low-value properties (less than 60,000). However, when dealing with highly valuable properties, the Random Forest is clearly the superior alternative, albeit even in these circumstances, it is recommended to treat its forecasts as a guide rather than an exact reality.

In conclusion, while our models are effective tools for projecting real estate values, residual analysis reminds us to utilise them with an understanding of their limitations and strengths. This allows us to take advantage of the greatest features of each model while remaining aware of when we should exercise caution with its forecasts.

A graph with different colored squares

Description automatically generated

# Figure 15: “Final visualization of model performance (Generated in Python using Matplotlib,2024).”

The performance chart provides an interesting comparison of our three prediction models. Two metrics are used to evaluate each model: R², which measures accuracy, and RMSE, which measures error.

Random Forest stands out as the most effective model, with the greatest blue bar and the lowest green bar, indicating that it delivers very accurate predictions with minimum errors. Linear Regression has a great performance, with good accuracy and moderate errors, making it a dependable option.

Finally, the SVR model has the lowest performance. Although it achieves an adequate level of accuracy (blue bar), the large green bar suggests that its forecasts contain much more errors. This shows that, among the three approaches, SVR is the least recommended for this particular prediction instance.

This visual comparison allows us to conclude unequivocally that Random Forest is the best fit for our dataset, followed by Linear Regression, although SVR would require significant modifications to be competitive.

### Data Sources

The database that will be worked on focuses on the information that we were recommended to take from the university website in the Strategic Thinking course. The teacher shared a dataset with different links and one of them is Kaggle. In fact, once on the page, we searched the building databases and found that this is one of the most comprehensive and has a wide variety of information.

Housing price of Beijing from 2011 to 2017, fetching from Lianjia.com

A screenshot of a website

Description automatically generated

### Ethical Considerations

The project of future housing relocations in the city of Beijing does not have any problem associated with the data that is being taken since all the information will be taken from this page: Beijing housing price from 2011 to 2017, obtaining from Lianjia.com. as far as it has been observed its database is public and this information does not contain confidential data, nor user privacy, in terms of social impacts it would be a good tool to apply in society but it would not be a tool to generate panic in society.

# Conclusion

To summarise, this detailed examination of Beijing's real estate market yielded substantial and enlightening conclusions. The Random Forest model consistently outperforms alternatives, with a R² of 0.9948 and the lowest normalised RMSE (0.25). This outstanding performance is more than just a technical feat; it has significant practical ramifications for all real estate market participants.

The model's resilience is demonstrated by its capacity to handle various price ranges and property features, but we have discovered that even this superior model requires extra attention when predicting high-value properties. The comparison with Linear Regression (R² = 0.8741) and SVR (R² = 0.8089) strengthened our trust in Random Forest as the main model, while providing valid backup options for various circumstances.

# References

<https://github.com/CCT-Dublin/capstone-project-feb-2024-ft-derlyai.git>

<https://github.com/derlyai/CA-2-Capstone-Report-Strategic-Thinking>

Housing price of Beijing from 2011 to 2017, fetching from https://www.kaggle.com/datasets/ruiqurm/lianjia

Beijing second-hand house Beijing rent Beijing real estate network Beijing Lianjia network. (2024). Recovered from: https://bj.lianjia.com/chengjiao

Clostermann, Zhong, Zhao, Li, Cheng, Ding. (2023). Capital Square Beijing Renovation. ARQA. Recovered from: https://arqa.com/en/architecture/capital-square-beijing-renovation.html [March 25,2024]

Zhicheng. (2020). Chinese growth of 6.1%, the lowest in 30 years. PIME Asianews. Recovered from: https://www.asianews.it/noticias-es/El-crecimiento-chino-del-6,1,-el-m%C3%A1s-bajo-en-30-a%C3%B1os-49052.html.

EURE (Santiago) vol.37 no.111 Santiago (mayo 2011). Recovered from: https://www.scielo.cl/scielo.php?pid=S0250-71612011000200010&script=sci\_arttext&tlng=pt.

Alberca, A. S. (2020, October 4). La librería Matplotlib | Aprende con Alf. Aprende Con Alf. https://aprendeconalf.es/docencia/python/manual/matplotlib/

Smith, P. (2019). Living in Dublin, 3rd ed. Dublin: Longman.[image-4.png](attachment:image-4.png)

Shrewsbury, M. (2011). The similarities in humans and non human primates, Journal of Anatomy, vol. 202, no. 4, p.51-59.[image-5.png](attachment:image-5.png)

ML | Handling Imbalanced Data with SMOTE and near Miss Algorithm in Python. GeeksforGeeks, 28 June 2019, www.geeksforgeeks.org/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/. Accessed 12 Aug. 2022

[image.png](attachment:image.png)

Müller, Andreas C, and Sarah Guido. Introduction to Machine Learning with Python : A Guide for Data Scientists. Beijing, O’reilly, 2017[image-2.png](attachment:image-2.png).

Vaughan, Daniel. Analytical Skills for AI et Data Science : Building Skills for an AI-Driven Enterprise. Beijing ; Boston ; Farnham ; Sebastopol ; Tokyo O’reilly Media, 21 May 2020.[image-3.png](attachment:image-3.png)

Solving Linear Regression in Python.GeeksforGeeks,2020 July 16,https://www.geeksforgeeks.org/solving-linear-regression-in-python/ . Accessed 16 May. 2024.

Bobbitt, Z. (2022, May 11). A gentle guide to sum of squares: SST, SSR, SSE. Statology. <https://www.statology.org/sst-ssr-sse/>

Bobbitt, Z. (2020, February 27). How to calculate mean squared Error (MSE) in Excel. Statology. https://www.statology.org/how-to-calculate-mean-squared-error-mse-in-excel/

OpenAI. (2024). ChatGPT (May 16 version) [How to interpret measures of central tendency].https://chat.openai.com/chat (https://chat.openai.com/chat)

Christoph Helma, Eva Gottmann, Stefan Kramer, Knowledge discovery and data mining in toxicology, Stat. Methods Med. Res. 9 (4) (2000) 329–358.

I.-N. Lee, S.-C. Liao, M. Embrechts, Data mining techniques applied to medical information, Med. Inf. Internet Med. 25 (2) (2000) 81–102.

Wu, J., Gyourko, J., & Deng, Y. (2016). Evaluating conditions in major Chinese housing markets. Regional Science and Urban Economics, 58, 12-25.

Li, V. J., Cheng, A. W. W., & Cheong, T. S. (2017). Home purchase restriction and housing price: A distribution dynamics analysis. Regional Science and Urban Economics, 67, 1-10.

Zhang, L., & Yi, Y. (2018). What drives housing markets: Fundamentals or bubbles? The Journal of Real Estate Finance and Economics, 56(3), 369-391.

Géron, A. (2022). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (3rd ed.). O'Reilly Media.

VanderPlas, J. (2023). Python Data Science Handbook: Essential Tools for Working with Data (2nd ed.). O'Reilly Media.

McKinney, W. (2022). Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter (3rd ed.). O'Reilly Media.

Wickham, H., & Grolemund, G. (2023). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data (2nd ed.). O'Reilly Media.

Chen, D., & Chen, C. (2021). Seaborn: Statistical Data Visualization. Journal of Open Source Software, 6(60), 3021.

Waskom, M. L. (2021). Seaborn: Statistical data visualization. Journal of Open Source Software, 6(60), 3021.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in R (2nd ed.). Springer.

Kuhn, M., & Johnson, K. (2019). Feature Engineering and Selection: A Practical Approach for Predictive Models. Chapman and Hall/CRC.

Raschka, S., & Mirjalili, V. (2023). Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow (4th ed.). Packt Publishing.

Fang, H., Gu, Q., Xiong, W., & Zhou, L. A. (2016). Demystifying the Chinese housing boom. NBER Macroeconomics Annual, 30(1), 105-166.

Sun, W., Zheng, S., Geltner, D. M., & Wang, R. (2017). The housing market effects of local home purchase restrictions: Evidence from Beijing. The Journal of Real Estate Finance and Economics, 55(3), 288-312.

Wang, Z., & Zhang, Q. (2014). Fundamental factors in the housing markets of China. Journal of Housing Economics, 25, 53-61.

García, S., Luengo, J., & Herrera, F. (2015). Data Preprocessing in Data Mining. Springer International Publishing.

Brownlee, J. (2020). Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python. Machine Learning Mastery.

Reis, M. S., & Braatz, R. D. (2021). Data preprocessing. In Process Monitoring and Data Analysis Methods (pp. 47-80). De Gruyter.

EURE (Santiago) vol.37 no.111 Santiago (mayo 2011). Recovered from: <https://www.scielo.cl/scielo.php?pid=S0250-71612011000200010&script=sci_arttext&tlng=pt>.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), e1301.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. The Journal of Machine Learning Research, 13(1), 281-305.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to Linear Regression Analysis (6th ed.). Wiley.

Sheather, S. (2009). A Modern Approach to Regression with R. Springer.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics Surveys, 4, 40-79.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications, 42(6), 2928-2934.

Mutanga, S. S., Ayanshola, A. M., & Anifowose, A. Y. (2022). A comparative analysis of machine learning algorithms for real estate price prediction. Journal of Big Data, 9(1), 1-20.