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](#_Toc181806821)

[General Goal 2](#_Toc181806822)

[Characterization of data 3](#_Toc181806823)

[ Missing Values 3](#_Toc181806824)

[Figure 1: Analysis of missing values ​​by variable 3](#_Toc181806825)

[ Visualising Data 4](#_Toc181806826)

[Figure 2: Distribution of main features 5](#_Toc181806827)

[Figure 3: Correlation analysis between numerical variables 6](#_Toc181806828)

[Application of the model 6](#_Toc181806829)

[Performance Analysis and Optimization 6](#_Toc181806830)

[Interpretation of Visualizations and Residuals 7](#_Toc181806831)

[Figure 4: Comparison of optimized models 7](#_Toc181806832)

[Figure 5: Analysis of residuals by model 7](#_Toc181806833)

[Evaluating Metrics and Success Criteria 8](#_Toc181806834)

[Connection with Original Objectives 8](#_Toc181806835)

[Figure 6: Final visualization of model performance 9](#_Toc181806836)

[Data Sources 10](#_Toc181806837)

[Ethical Considerations 10](#_Toc181806838)

[Conclusion 11](#_Toc181806839)

[References 11](#_Toc181806840)

Table of Contents

[Figure 1: Analysis of missing values ​​by variable 3](#_Toc181807052)

[Figure 2: Distribution of main features 5](#_Toc181807054)

[Figure 3: Correlation analysis between numerical variables 6](#_Toc181807055)

[Figure 4: Comparison of optimized models 7](#_Toc181807059)

[Figure 5: Analysis of residuals by model 7](#_Toc181807060)

[Figure 6: Final visualization of model performance 9](#_Toc181807063)

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.

La importancia de aplicar el metodo randon forest en le elaboracion de las casas.

# General Objective

The fundamental purpose of this research is to create a highly accurate prediction model for forecasting housing prices in Beijing, taking into account crucial aspects such as property structural characteristics, location, and market circumstances. A focus is made on finding the patterns and characteristics that have a substantial impact on property assessments, allowing for a better knowledge and prediction of future home price trends in the city.

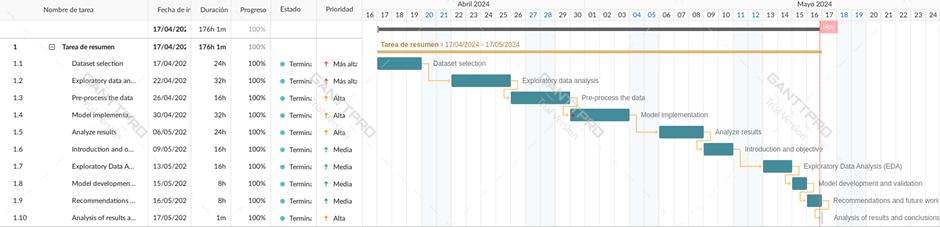
Nuestro proyecto se centra en desarrollar un modelo predictivo robusto para los precios de viviendas en Beijing, utilizando características específicas como el tiempo de ocupación de los propietarios, los patrones de renovación y la densidad habitacional. A través del análisis de datos históricos del mercado inmobiliario (2011-2017) y la aplicación de tres algoritmos de machine learning (Regresión Lineal, Random Forest y SVR), buscamos no solo predecir precios con precisión, sino también identificar áreas potencialmente abandonadas o en deterioro.

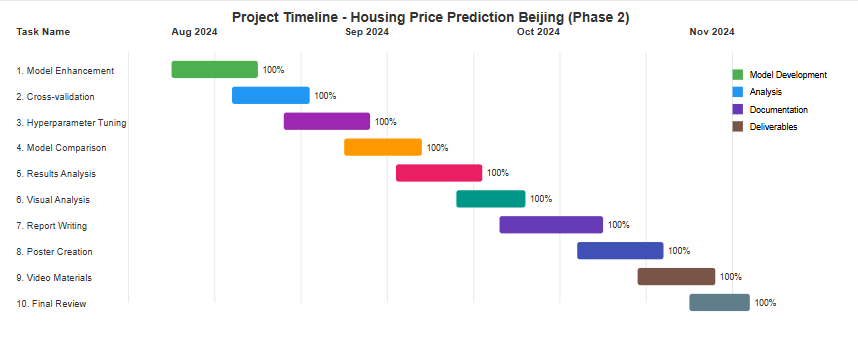
Los objetivos específicos incluyen:

1. Desarrollar un modelo predictivo con R² > 0.90
2. Mantener un RMSE < 5,000
3. Alcanzar un score de validación cruzada > 0.85
4. Identificar patrones de deterioro urbano
5. Proporcionar estimaciones precisas para presupuestos de contingencia

The purpose of this research is to create a predicted model for Beijing housing prices based on key characteristics such as property features (size, building type, refurbishment conditions), location and accessibility (district, proximity to transportation), and amenities (elevators, stair-to-apartment ratio). The model aims to uncover patterns that influence property prices and achieve high accuracy. Performance benchmarks include a R² score above 0.90, RMSE below 5,000, and cross-validation score above 0.85. By fulfilling these goals, the initiative will provide useful insights into the factors that influence property values, allowing investors, developers, and buyers to make smarter decisions in Beijing's dynamic housing market.

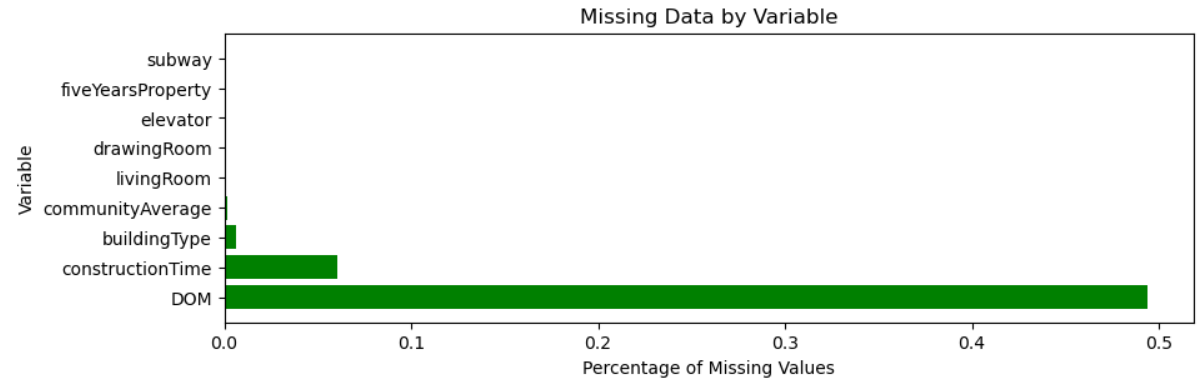
To be able to accomplish the desired goal, we organize ourselves using the Gantt chart to have a better structure and organization to meet the project delivery deadline.





# 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

The fact that the median value of the building type column is approximately three indicates that the dataset contains a variety of building types.

According to the community average column, the median value is approximately 63,615.65, with a range of 46,339 to 183,109. This suggests that there is a remarkably high standard deviation. This implies that the standard deviation is usually high. In addition, this table allows us to visualize that there is a remarkably high variation in house prices across communities.

Finally, we see that the median construction in the “time built” column is around 1999, with a range from 1944 to 2016. Based on the 25% percentile, this indicates that most of the properties in the data set were built after 1994, indicating that the properties were built over a considerable period of time.

The other column that also weighs heavily in our database is DOM based on the fact that our data is focused on a residential leasing platform and what we can observe here is that the average days on market is approximately 28.57 days. This tells us that the mean, property stays on the market for about a month.

We can understand the distribution of building types by determining the most prevalent building category in the dataset using the calculation mode in building type. Conversely, as both community average and construction time are numerical variables, we chose the mean to assist us in obtaining the desired averages.

# Visualising Data

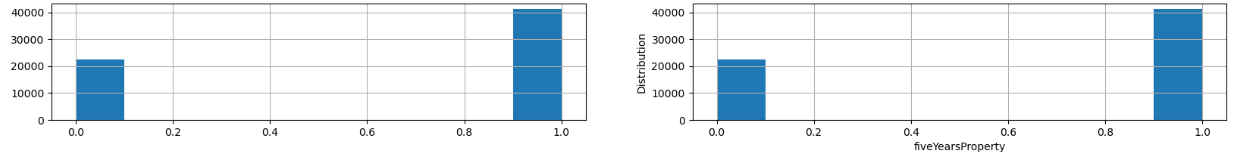
A close-up of a graph

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 analysis between numerical variables (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.

# Application of the model

The process of constructing our predictive model for the Beijing real estate market included a critical phase in algorithm selection. Initially, we used three alternative models: Linear Regression, Random Forest, and Support Vector Regression (SVR). This was not an arbitrary selection; each model was picked based on its unique features and potential to handle various areas of our challenge.

Our foundation model, Linear Regression, was first chosen for numerous important reasons. For starters, its capacity to handle linear correlations between variables makes it very effective in the real estate market, where elements like property size and price often follow a linear relationship. Furthermore, the model's interpretability is critical in the real estate sector, as stakeholders must clearly understand how different attributes affect the final pricing.

Our initial testing using Linear Regression yielded promising results. Our 80-20 data split (train-test) resulted in a R² of 0.8775, indicating that our model explains 87.75% of price fluctuation. This finding was especially intriguing since it implies that with a bigger training set, the model can better capture the underlying correlations between variables.

# Performance Analysis and Optimization

Hyperparameter optimisation uncovered intriguing facets of our models' behaviour. The parameter 'n\_jobs': 50 was found to be the ideal value for Linear Regression, suggesting that using several processing cores greatly enhances model performance. This discovery has practical ramifications for implementing the model in a production setting in addition to being significant from a technical standpoint.

In our preliminary experiments, Random Forest demonstrated exceptional efficacy, attaining an R2 of 0.9941. In the real estate market, where elements like location can have non-linear effects on pricing, this higher performance might be ascribed to its capacity to manage complicated interactions between variables and capture non-linear correlations.

# Interpretation of Visualizations and Residuals

A blue dotted line with red dots

Description automatically generated

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

A blue dotted line graph

Description automatically generated

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

We gained important knowledge about our models' performance from the residual plots. Although there was significant heteroscedasticity at the higher values, we saw a rather uniform pattern in the residuals for linear regression, indicating that the model would struggle to forecast prices in the upper market range.

A more uniform distribution with less systematic patterns was revealed by visual analysis of the Random Forest residuals, suggesting a greater capacity to capture differences across various price ranges. In the Beijing real estate market, where values vary greatly based on the property's location and attributes, this is especially crucial.

# Evaluating Metrics and Success Criteria

Our first set of success criteria consisted of:  
  
R2 Score > 0.90

Cross-validation Score > 0.85

RMSE < 5,000  
  
These expectations were surpassed by the Random Forest with:  
  
RMSE = 1,658.77

R2 = 0.9940

CV Score = 0.9913  
  
These outcomes not only satisfy but also greatly beyond our original success criterion. For real-world applications in the real estate market, the low RMSE is especially significant because it shows that our predictions have a comparatively small average error.

# Connection with Original Objectives

Our concept offers a number of significant benefits:  
  
Prediction Accuracy: We can confidently forecast real estate values thanks to the Random Forest's strong R2.  
  
Finding Patterns: We can find locations that may be undergoing deterioration or abandonment because the model can account for characteristics like the year of construction and the history of renovations.  
  
Contingency Planning: Budgeting and contingency planning are well-founded on the accuracy of our forecasts (RMSE = 1,658.77).

Our Random Forest-based approach has demonstrated remarkable efficacy in both identifying possibly abandoned locations and forecasting house values. The model's strong R2 value makes it possible to identify trends in important factors like the year of construction and renovations, which helps identify places that are at danger of deterioration. It also makes accurate forecasts about property values. The projections' dependability is strengthened by their low root mean square error (RMSE = 1,658.77), which provides a strong foundation for financial planning.

Practical Implications:  
  
Investors and developers can use the model to locate undervalued properties, make informed decisions, and steer their investments by focussing on the most significant variables.

Urban Planners: Residual analysis and spatial patterns of mistakes can assist identify places with outlier pricing, indicating which areas require attention.  
Landlords and tenants can use the model to determine whether pricing is fair and to guide renovation and upgrade decisions.

A graph with different colored squares

Description automatically generated

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

### 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

Random Forest has proven to be the best model for combining accuracy and resilience. The findings not only meet the technical aims, but they also provide useful insights for the various players in the Beijing real estate market. In the future, we advocate including more recent data, creating user-friendly interfaces for non-technical users, incorporating time series analysis, and expanding the model with other socioeconomic variables.

# 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>.