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](#_Toc181725263)

[General Goal 2](#_Toc181725264)

[Characterization of data 3](#_Toc181725265)

[ Missing Values 3](#_Toc181725266)

[ Visualising Data 4](#_Toc181725267)

[Application of the model 6](#_Toc181725268)

[Hyperparameter 6](#_Toc181725269)

[Data Sources 7](#_Toc181725270)

[Ethical Considerations 8](#_Toc181725271)

[Conclusion 9](#_Toc181725272)

[References 10](#_Toc181725273)

Projection of future housing relocations in Beijing

# Introduction

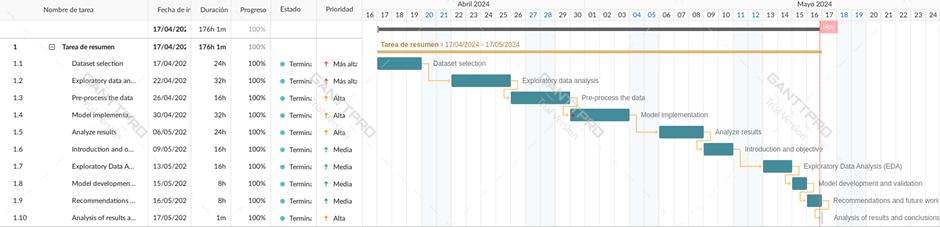
Beijing, China's capital city, has undergone tremendous urbanization in recent decades. A significant example is CLOU's Capital Square Beijing makeover project, which attempts to modernize and reposition public places using the "City Lantern" idea. This urban revitalization project, combined with broader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

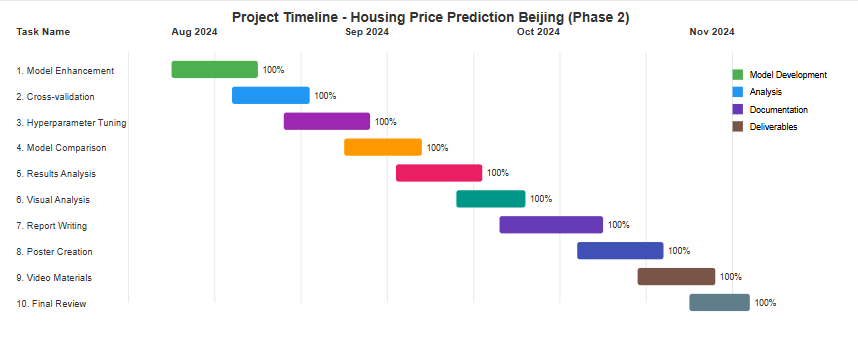
In 2019, the Chinese economy's growth rate decreased to 6.1%, the lowest in 30 years, despite GDP per capita topping $10,000 for the first time. This economic framework, which includes decreasing infrastructure investment (from 4% to 3.8%) and real estate industry investment (from 10.2% to 9.9%), is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

# General Goal

The primary goal of this study is to create a forecast model for housing prices in Beijing utilising specific factors such as owner occupation time, remodelling patterns, and housing density. We hope to not only accurately predict prices but also identify potentially abandoned or deteriorating areas by analysing historical real estate market data (2011-2017) and applying three machine learning algorithms (Linear Regression, Random Forest, and SVR) based on the year of construction and renovation. This study will allow for more accurate contingency budget estimates that take into account critical elements such as location, property size, and remodelling patterns, resulting in better urban planning and decision-making in Beijing's real estate market.

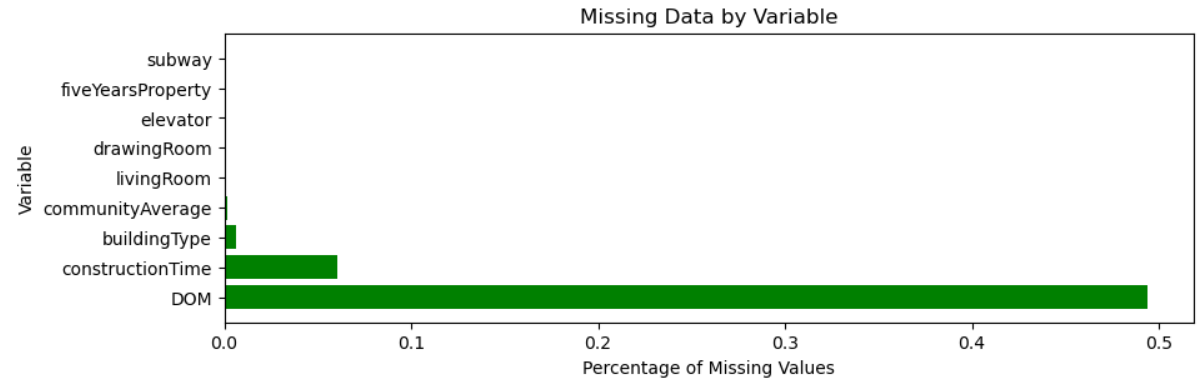
To be able to accomplish the desired goal, we organize ourselves using the Gantt chart to have a better structure and organization to meet the project delivery deadline.





# Characterization of data

# Missing Values

****

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

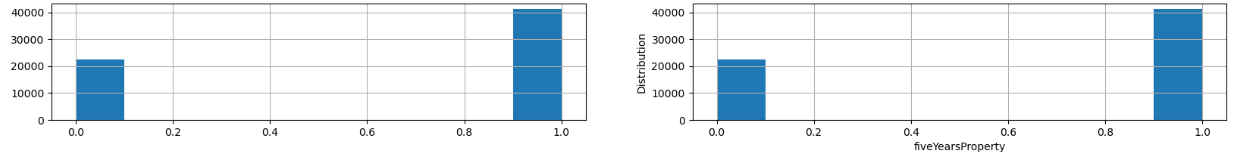
A close-up of a graph

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

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

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

It is noted that the choice of Linear Regression is mainly due to its ability to efficiently handle diverse and complex data sets, which is typical in the real estate sector, where variables such as location, property size and year of construction significantly influence pricing strategies.

Linear Regression

R2 score for LR on 0.2: 0.877532530096987

R2 score for LR on 0.25: 0.6676743926861246

R2 score for LR on 0.3: 0.6894598161216585

With this information provided by the linear regression model we can see that when the R² of 20% of the data is applied as the test set (R² = 0.8775), which suggests to us that with a larger training set (80%), the model can better capture the relationship between the variables.

# Hyperparameter

Based on the results of fitting a linear regression model with multiple training data splits and hyperparameter optimization we can conclude that the best parameters for linear regression are {'n jobs': 50}, indicating that the use of multiple processing cores helps to improve the model performance.

Second, the best R2 value obtained is 0.8708, indicating that the model explains approximately 87.08% of the variability of the target variable.

Thirdly, the RMSE (relative cubic error) is 7823.33, providing us with a measure of the average deviation between the predicted values and the actual values.

In addition, the intercept of the model is -7648042.55, and the coefficients associated with the characteristics indicate the magnitude and direction of the impact of each characteristic on the target variable.

### 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

The investigation reveals that the random forest regression model outperforms the linear regression model in terms of prediction accuracy. The Mean Squared Error (MSE) of the random forest model is 3,634,563.68, a significant decrease from the MSE of 61,204,567.36 for the linear regression.

With an R-squared of 0.9923, the variability of the target variable is explained in 99.23% of cases. Additionally, the scatter plot visualisation shows that the random forest maintains great precision throughout the target variable's complete range, in contrast to the linear regression, which loses precision at higher values.   
  
The linear regression model performs rather well and has easily interpreted coefficients, but its non-normal distribution of residuals and decreasing accuracy for higher target values limit its use.

Going ahead, the knowledge acquired by assessing the performance of these models through the use of statistical tests and quantitative measures will direct the choice of suitable approaches for further predictive modelling initiatives. To guarantee the most precise predictions, it will be crucial to keep refining hyperparameters, addressing data distribution assumptions, and putting sophisticated preprocessing and feature engineering into practice. All in all, this analysis shows how important it is to choose models carefully and to evaluate them thoroughly in order to facilitate effective data-driven decision-making.

Finally, in order to assess the project's possibilities, threats, weaknesses, and strengths, a comprehensive SWOT analysis would be helpful. Leveraging crucial assets such as the comprehensive data set on home attributes and pricing, along with advanced modelling approaches like linear regression, is possible. On the other hand, there are vulnerabilities and risks that need to be handled, such as possible data gaps, difficulties in accounting for all the variables that affect prices, and uncertainty about upcoming market developments. Opportunities exist to monetize prediction models, extend study to other locations, and use insights to inform urban development. Effective housing data collection, analysis, and utilisation to produce insightful business knowledge is the cornerstone of this project's success. To maximise the benefits of this data-driven initiative and guide strategic decision-making regarding urban renewal and real estate market dynamics, it will be imperative to ensure the quality of the data, choose appropriate analytical techniques, translate findings into actionable recommendations, and update the models on a regular basis.

# References

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.