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 2 Capstone Report |
| *Assessment Due Date* | 16 April 2024 |
| *Date of Submission* | 19 May 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](#_Toc166871603)

[Objective 2](#_Toc166871604)

[Characterization of data 3](#_Toc166871605)

[ Missing Values 3](#_Toc166871606)

[ Visualising Data 5](#_Toc166871607)

[Application of the model 6](#_Toc166871608)

[Hyperparameter 7](#_Toc166871609)

[Data Sources 7](#_Toc166871610)

[Ethical Considerations 8](#_Toc166871611)

[Conclusion 9](#_Toc166871612)

[References 9](#_Toc166871613)

Projection of future housing relocations in Beijing

# Introduction

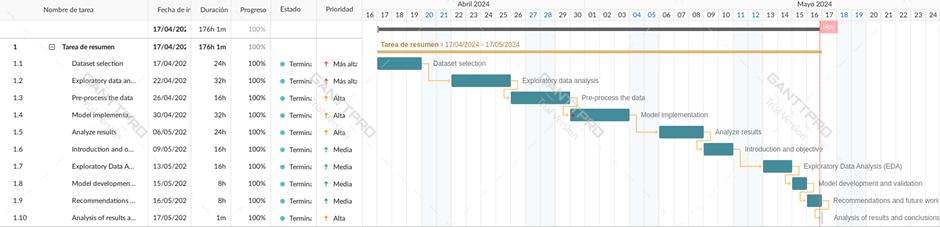
In this project, one of the most evident urban phenomena in the last decades in Beijing is being detected, this is the process of urban renewal of the city center, which CLOU has been managing to renovate inside the Capital Square Beijing and what they want to achieve with this is to reposition and improve the public space.

What is intended with this project is to generate a noticeable impact on the real decay and abandonment that is being observed in houses and apartments. Additionally, in these data it is observed that some houses were built since 1960. Therefore, it is intended to analyse the number of times that these houses have been remodelled to date and the variation of prices that have been obtained according to the structure of the building.

The objective of this project is to be able to predict the price of housing based on certain specific characteristics, such as whether the owner has been in the building for more or less than 5 years, the number of people who are in each building, it is considered that with this information it can be determined which areas of the perimeter are abandoned based on the year of construction and the number of renovations that have been made since the time the house was built. To do this, the regression algorithm will be used to determine how prices vary according to the parameters mentioned above. with this information, a percentage of money for contingencies that may arise over time could be contemplated. However, for this budget to be as tight as possible, it should be taken into account that it will vary depending on the location of the place, the size of the house, the number of people living in the place and the amount of remodeling that has been done or will be done over time.

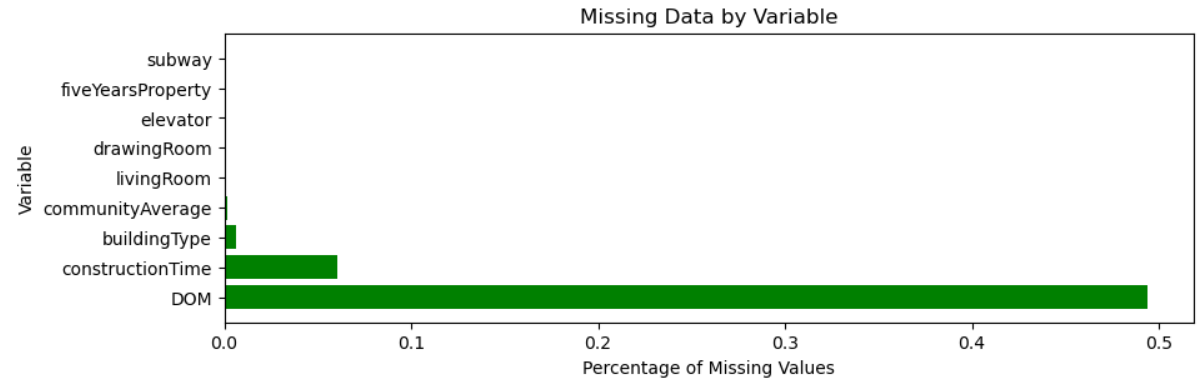
To obtain the results of the above, we have relied on the use of linear regression, root mean square error (RMSE) and coefficient of determination (R^2). Simultaneously the main objective is to predict the price of housing from certain characteristics such as house size, location, number of bedrooms among others. These models will allow me to build, evaluate and adapt the results to obtain the most accurate predictions and better understand the relationships between the characteristics of the houses and their prices.

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 64,330.59, with a range of 20,483 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.

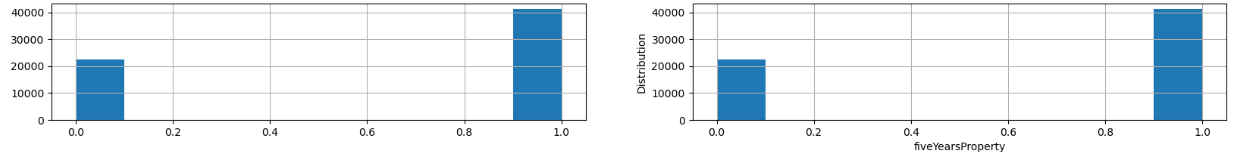
A close-up of a graph

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.

# Visualising Data

**Histogram of price**

A graph of a graph of a number of blue bars

Description automatically generated with medium confidence

We can see the association between the numerical variables after cleaning our database. To illustrate how it affects the different types of variables such as construction time, renovation condition in relation to our target variable price.

A screenshot of a computer screen

Description automatically generated

The correlation graph between numerical variables shows that the variables “living room”, “drawing room”, “elevator” and “building structure” among others, are represented by yellow dots approaching the central yellow line. This suggests that these variables have a moderate correlation with each other and with the central trend 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 analysis shows that in terms of prediction accuracy, the random forest regression model performs better than the linear regression model. The random forest model's Mean Squared Error (MSE), which is 3,634,563.68, is substantially lower than the linear regression's MSE of 61,204,567.36.

The target variable's variability is explained in 99.23% of cases, with an R-squared of 0.9923. Furthermore, the scatter plot visualisation demonstrates that, in contrast to the linear regression, which loses precision at higher values, the random forest retains strong precision over the whole range of the target variable.   
  
The non-normal distribution of the residuals and diminishing accuracy for higher target values restrict the employment of the linear regression model, despite its reasonable performance and interpretable coefficients.

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.

# References

<https://github.com/derlyai/Machine-Learning--10-ETCS--.git>

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.