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](#_Toc165057900)

[Objective 2](#_Toc165057901)

[Characterization of data 3](#_Toc165057902)

[ Missing Values 3](#_Toc165057903)

[ Followed prices associated to the building type 4](#_Toc165057904)

[Hyperparameter 6](#_Toc165057905)

[Conclusion 6](#_Toc165057906)

[References 7](#_Toc165057907)

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 analyze the number of times that these houses have been remodeled to date and the variation of prices that have been obtained according to the structure of the building.

A screenshot of a project

Description automatically generated

# Objective

The objective of this project is to be able to predict the price of housing based on certain specific characteristics, such as the type of building, the structure of the premises and whether the owner has been on the site for more or less than 5 years, by determining which areas of the perimeter are abandoned based on the year of construction and the number of renovations. To do this, sorting algorithms are used that classify perimeter areas into categories of abandonment. For example, there are areas with high abandonment, areas with moderate abandonment and areas with no abandonment. This will allow us to know the estimated cost of the changes that have been made to the property. Then, with this information, a percentage of money could be contemplated for contingencies that may arise over time. 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 has been done to date.

El objetivo de este proyecto es poder predecir el precio de la vivienda en función de ciertas características específicas, como el saber si el propietario lleva en él edificio más o menos de 5 años, la cantidad de personas que están en cada edificacion, se considera que con esta información se puede determinar qué zonas del perímetro están abandonadas en función del año de construcción y el número de renovaciones que se han realizado desde el momento en que se construyó la vivienda. Para ello, se utilizara el algoritmo de regresión para poder determinar cómo varían los precios según los parámetros mencionados anteriormente. con esta información, se podría contemplar un porcentaje de dinero para imprevistos que puedan surgir con el tiempo. Sin embargo, para que este presupuesto sea lo más ajustado posible, se debe tener en cuenta que variará dependiendo de la ubicación del lugar, el tamaño de la casa, el número de personas que viven en el lugar y la cantidad de remodelaciones que se hayan hecho o se piensan realizar con el paso del tiempo.

To obtain the results of the above, we have relied on the use of linear regression and Random Forest. Simultaneously the main objective is to predict the price of housing from certain characteristics such as house size, location, number of bedrooms among others. What was done was to train the model using those characteristics. As for the model, it will be evaluated using metrics such as (R2) in Linear Regression since this method will allow us to evaluate the performance of the model and in Random Forest.

# Characterization of data

# Missing Values

**A green bar graph with white text

Description automatically generated**

The distribution of the data in the "buildingType", "communityAverage" and "constructionTime" 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 text

Description automatically generated

The fact that the median value in the "Building type" column is approximately three indicates that the data set contains a variety of building types.

According to the "communityAverage" 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. Also, This suggests that there is 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 dataset were built after 1994, indicating that the properties were built over a considerable period of time.

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

A close-up of a graph

Description automatically generated

In the plot we can see that by applying the Mean in the Community Average column, some artificial values are introduced into the original data distribution, which alters its shape and results in a more unbiased and marginally left-skewed normal probability density curve.

# Followed prices associated to the building type

We can see here the association between the categorical variables after cleaning up our database. To illustrate how the variables, affect the several building types tower, bungalow, plate/tower, and plate in relation to our objective variable price, we created a box diagram. It is evident from this graph that the bungalow-style buildings are more expensive than the norm.

**Histogram of price**

A graph of a graph of a number of blue bars

Description automatically generated with medium confidence

It can be seen that the frequency of viewing ceases at RMB 150,000 after slowly stabilizing at a higher cost. house prices are analyzed and predicted using the linear regression model, which provides crucial information on the various elements that are key to real estate values. Given the complexity of the real estate market and the existence of both linear and non-linear correlations among the many variables that influence housing prices, this methodology is particularly appropriate.

The choice of Linear Regression is mainly due to its ability to efficiently handle diverse and complex data sets, which is typical in real estate, where variables such as location, property size and age significantly influence pricing strategies.On the other hand, the Random Forest trick allows us to operate in high-dimensional spaces, making it an ideal tool for our data set, which includes a variety of both numerical and categorical characteristics.

Linear Regression

R2 score for LR on 0.2: 0.9094735990400029

R2 score for LR on 0.25: 0.9090839513276163

R2 score for LR on 0.3: 0.908099775346926

Random Forest

R2 score for 0.2 on 0.9788752811130561

R2 score for 0.25 on 0.9778537599826562

R2 score for 0.3 on 0.9761005667074901

The techniques employed for the variation in precision in three training divisions (20%, 25%, and 30%) were linear regression and random forest. Cross-validation was used to assess these models' performance across various data divisions. The results showed that in every division, the random forest outperformed the linear regression with higher R2 values in each case.

According to the prior data, linear regression accounts for 90.9% of the variation in the price variable across the various data divisions. On the other hand, 97.0% of the variability of the price variable in the different data splits are explained by the random forest, which indicates that it is the most efficient method to represent the relationship between the characteristics and the target variable in this particular case.

# Hyperparameter

The GridSearchCV technique allows me to evaluate the performance of each combination of hyperparameters and ultimately select the most optimal one.

GridSearchCV is applied to tune the hyperparameters in linear regression model and random forest model. For the linear regression model, the "n\_jobs" hyperparameter is tuned using a grid of predefined values. Cross validation is used to evaluate the performance of the model for each value of "n\_jobs" and the value that produces the best performance is selected. Similarly, it is done for the random forest model, multiple hyperparameters such as “n\_estimators”, “max\_ Depth”, “min\_samples\_split”, “min\_samples\_leaf” and “max\_features” are tuned. GridSearchCV is again used to find the optimal combination of these hyperparameters.

At the end of the process, the best hyperparameters found for each model are obtained, along with the performance associated with the use of the specific evaluation metric (in this case, the score R2 for linear regression and the average score R2 for random forest). These optimized hyperparameters are further used to train final models to maximize their performance on new data sets.

# Conclusion

this project is completed, it can be concluded that the phenomenon of the curse of dimensionality was experienced in the development of the project. This fact according to Richard E. Bellman refers to a small amount of data compared to the intrinsic dimension of the data. For example: in a three-dimensional space as when we look through a telescope; We can easily visualize small particle dots in a cube. However, when we do the Hyperparameter percentages were very good with 98% but according to the data and analysis it could be observed that the data is overfitting.

This situation affects my data because I want to obtain reliable results. On the other hand, the amount of data needed would often continue to grow exponentially with dimensionality, which would make the analysis more tiring and take longer to analyze. Finally, when organizing and searching for high-dimensional data, objects would appear scattered in different ways, making common organization and search strategies difficult. In general, processes carried out in machine learning, classification and prediction models could fail due to lack of sufficient data in high-dimensional spaces.

# 

# 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.