

**Pper on the use of machine learning techniques to identify the fundamental and macroeconomic factors affecting property prices and property price changes in Hong Kong**

**a dissertation submitted in partial fulfilment of the   
requirements for a****Master of Science (MSc)**

**In Artificial Intelligence**

**at the University of Hull**

**by**

**Student No: 202249387**

**9 January 2025**

**COPYRIGHT STATEMENT**

Copyright in of this MSc paper belongs to the author Mr Tim Hallworth, please contact him at T.HALLWORTH-2022@hull.ac.ukif you wish to use any paper of this paper

Dedication

I dedicate this paper to my beloved wife, Ms. Fenny Tsang, whose unwavering support and patience have been the cornerstone of my research journey. Her encouragement and understanding allowed me the time and space to delve deeply into my work, while her willingness to be a sounding board for my frustrations and wild theories provided me with the clarity and motivation needed to continue.

Thank you, Fenny, for your endless love, support, and belief in me. This work would not have been possible without you by my side.

Acknowledgements

I would like to express my deepest gratitude to my esteemed professor, [Professor XXX], for his unwavering support, guidance, and inspiration throughout the course of this research. His mentorship has been invaluable in shaping the direction and quality of this work.

I am also profoundly thankful to Mr. Parakrant Sarkar and Mr. Michel Leary, whose thorough reviews and insightful suggestions significantly enhanced the clarity and depth of this paper. Their contributions have been instrumental in refining the ideas presented here.

This work is dedicated to all of you. Thank you for your encouragement and for helping me.

Abstract

This project used machine learning algorithms[[1]](#footnote-1) to predict residential property prices and price changes in Hong Kong.

The objective was to utilize artificial intelligence (AI) and existing property data to analyze the factors driving property prices and changes and predict future property price trends in Hong Kong.

The project used two databases. One was a large residential sales dataset prepared by one of the largest property agencies in Hong Kong. This comprised 159,676 entries containing property sales transactions from 2020 to 2023 from 18 geographical districts. [[2]](#footnote-2) A second dataset incorporated macroeconomic indicators based on the census and statistics department of the government of Hong Kong.  The two datasets were then merged.

The merged dataset was then reviewed and cleaned up before various machine learning algorithms were applied. This involved several steps: first, examining and understanding the data from the property databases, performing data analysis, cleaning, managing, and normalizing the dataset. Then, the merged dataset was analyzed using various machine-learning techniques, including regression, decision trees, XGBoost and MLP.

Additionally, hyperparameter optimization was carried out using these methodologies to enhance the accuracy of the results and predictions.

The machine learning algorithms identified the key factors and the macroeconomic factors affecting property price changes, but the datasets were incomplete, and improving the datasets would have been time-consuming and beyond the scope of the study.

Contents

[Dedication i](#_Toc176284372)

[Acknowledgements ii](#_Toc176284373)

[Abstract iii](#_Toc176284374)

[Contents iv](#_Toc176284375)

[List of Figures xiv](#_Toc176284376)

[List of Tables xv](#_Toc176284377)

[Definitions xvi](#_Toc176284378)

[Chapter 1 Introduction 1](#_Toc176284379)

[1.1 Factors affecting Property Prices 1](#_Toc176284380)

[1.2 Ethical and Copyright Issues [Placeholder] 2](#_Toc176284382)

[Chapter 2 Literature Review 4](#_Toc176284383)

[Chapter 3 Methodology 5](#_Toc176284384)

[3.1 Database Selection 5](#_Toc176284385)

[3.2 Review of the merged\_dataset 5](#_Toc176284390)

[3.3 Identification of targets and factors 6](#_Toc176284401)

[3.4 Data Examination 7](#_Toc176284405)

[3.5 Data Cleaning 7](#_Toc176284416)

[3.6 Data Cleaning Techniques Used 8](#_Toc176284417)

[3.7 Feature Engineering 9](#_Toc176284425)

[Chapter 4 Data Analysis Results 11](#_Toc176284429)

[4.1 Preliminary data analysis 11](#_Toc176284430)

[4.2 Checking the completeness of the data 13](#_Toc176284442)

[4.3 Correlation Analysis 14](#_Toc176284443)

[4.4 Multicollinearity Analysis 14](#_Toc176284449)

[Chapter 5 Machine Learning Results 16](#_Toc176284453)

[5.1 Regression Analysis 16](#_Toc176284454)

[5.2 Decision Trees Analysis 17](#_Toc176284459)

[5.3 XGBoost 17](#_Toc176284463)

[5.4 MLP 19](#_Toc176284470)

[Chapter 6 Discussion on results 20](#_Toc176284474)

[**6.1** **Summary of Results** 20](#_Toc176284475)

[**6.2** **Comparison of Machine Learning Results** 20](#_Toc176284478)

[6.3 Problems and Challenges 21](#_Toc176284479)

[Chapter 7 Conclusion 23](#_Toc176284480)

[7.1 Model Reliability 23](#_Toc176284481)

[7.2 Future Research and last thoughts 23](#_Toc176284482)

[Reference list / Bibliography 25](#_Toc176284483)

[Appendix 1 I](#_Toc176284484)

List of Figures

Figure 1 -

Figure 2 -

Figure 3 -

List of Tables

Table 1 – Comparison of XGBoost Results using tuned and untuned hyper parameter models

Table 2 - Comparison of Machine Learning Results

Table 3 -

Definitions

AI - Artificial Intelligence

HK – Hong Kong

ML - Machine Learning

MLP - Multilayer Preception

XGBoost - Extreme Gradient Boosting

# Introduction

From 2008 to 2013, residential property prices in Hong Kong skyrocketed by 134%. Despite a market slowdown in the first half of 2014, house prices then surged by 41.5% between the second half of 2016 and the first half of 2018.[[3]](#footnote-3)

This outperformed the stock markets in most countries. In Hong Kong, during this period, people who invested in residential property made significant returns on their investments.

The potential for property development is limited in Hong Kong, as with a population of 6.5 million people and a relatively small area of only 1,108 km² (much of which is hilly and mountainous), land is in short supply.[[4]](#footnote-4)

As a result, driven by low interest rates, China's growth, other factors, and confidence in the certainty of property ownership based on the English legal system, residential property prices in Hong Kong are among the highest in the world.[[5]](#footnote-5)

However, residential property prices have been falling since 2018. The question is has this era of growth ended, does Hong Kong property still offer opportunities for investors and what factors would affect future property price changes.

By utilizing AI, the project hoped to provide insight into these questions.

## Factors affecting Property Prices

### Size and Location

### Like all property worldwide, residential property prices are influenced by two fundamental factors: property location and the size of the property.[[6]](#footnote-6)

* + 1. Macroeconomic factors

However, property price changes are influenced by various macroeconomic factors, which research has identified. This project uses machine learning to identify the fundamental and macroeconomic factors that could affect property prices in Hong Kong.

The key macroeconomic factors considered in this study include:

#### CPI (Consumer Price Index): This reflects the changes in the price level of a basket of consumer goods and services, affecting overall affordability of property as a major purchase.

#### IR (Interest Rate): The cost of borrowing, can significantly affect property demand and prices.

#### MW (Monthly Wage): The average monthly wage levels impact buying power and the ability to afford property.

#### M3 (Money Supply M3): This represents the total money supply within the economy, influencing liquidity and spending power.

#### SD (Savings Deposit): This is the level of saving deposit, which reflects the availability of funds for property purchases.

#### UR (Unemployment Rate): A higher unemployment rate reduces the number of potential buyers, impacting property demand and prices.

#### GDP (Gross Domestic Product): This number reflects economic growth, which affects income levels and investment capacity.

#### CI (Confidence Index): This measures consumer sentiment, which impacts spending and investment decisions, including property purchases.

#### SM (Stock Market): The stock market performance can influence wealth and investment decisions, affecting property demand.

#### LTV (Loan to Value): Loan-to-value ratios impact borrowing capacity and, thus, property affordability.

#### HS (Housing Starts): The number of new housing projects that have started can affect supply and property prices.

#### SOLD (Number of Properties Sold): The volume of property transactions can indicate market activity and price trends.

#### PG (Population Growth): Population growth can drive the demand for housing and influence property prices.

#### CG (China Growth): Economic growth in China is very important to Hong Kong and can have indirect effects on the City's property market through investment flows.

#### PC (Property Price Change): The historical year-on-year property price changes can provide insights into future price trends.

The project aimed to use machine learning to identify the above fundamental and macroeconomic factors that could affect property prices in Hong Kong.

## Ethical and Copyright Issues [Placeholder]

As part of this research, a dataset from a national property agency based in Hong Kong, which was made publicly accessible via Kaggle, was utilized.

The dataset contains information on 160,000 property transactions in Hong Kong. While the transactional data was compiled by the property agency, the underlying data is publicly available and can be accessed through the Land Registry of the Government of Hong Kong. Additionally, macroeconomic data from the National Census and Statistics Department of the Government of Hong Kong, which is also publicly available, was incorporated into this study.

The dataset used does not contain any personal information or identifiable details of individuals, thereby minimizing privacy concerns and I have given careful consideration was given to the copyright and licensing conditions associated with the data.

Having determined that the data is within the public domain and not subject to specific licensing restrictions, it is concluded that no copyright issues exist regarding its use in this academic research.

In addition to the legal considerations, the ethical implications of employing this dataset were thoroughly evaluated. While the data is publicly accessible, the potential societal impacts of the research outcomes were carefully considered and I do not consider that the research lead to the automation of processes currently performed by humans, or that it would have significant implications for employment and broader societal dynamics. Although it is not possible to fully anticipate the broader effects of this research, a commitment has been made to remain cognizant of these implications and to apply the findings and any resulting technologies in an ethically responsible manner.

Furthermore, transparency regarding the source and application of the data has been maintained throughout the research process. This includes acknowledging the limitations of the data and exercising caution in the interpretation of the results. By adhering to these principles, this research seeks to contribute to the academic field in a manner that is both legally sound and ethically responsible.

# Literature Review

To effectively prepare for this project, an extensive literature review was undertaken to explore the a range of factors that influence residential property prices in Hong Kong.

This review involved a thorough examination of both academic papers and industry news sources, ensuring a broad and detailed understanding of the topic. The academic papers provided a theoretical framework, offering insights into the economic, social, and political factors that have historically affected property prices in Hong Kong.

These sources helped identify key variables such as interest rates, government policies, supply and demand dynamics, and global economic trends.

Additionally, the review of industry news sources offered a more current perspective, highlighting recent trends and developments within the Hong Kong real estate market. This included analysis of market sentiment, investor behavior, and the impact of recent events such as the COVID-19 pandemic and geopolitical tensions. By combining academic research with up-to-date industry data, the literature review provided a comprehensive overview of the multifaceted factors that contribute to residential property price fluctuations in Hong Kong.

This thorough understanding formed the foundational basis for the project, guiding the subsequent analysis and ensuring that the research was grounded in both theory and practice.

# Methodology

The project utilized two datasets: a large residential dataset of sales transactions from a major Hong Kong property agency and a macroeconomic dataset of indicators from the Hong Kong government's census and statistics department. These datasets were merged and underwent data examination, cleaning, management, and normalization.

The methodology involved applying machine learning algorithms, including regression,[6] decision trees,[[7]](#footnote-7)XGBoost,[[8]](#footnote-8) and multilayer perceptron (MLP),[[9]](#footnote-9) to the merged dataset. Hyperparameter optimization was also performed to enhance prediction accuracy.

## Database Selection

The project utilized two datasets and merged them into a single dataset called the merged\_dataset for analysis. The two datasets are as follows:

### Property Transactions Dataset

### This dataset contains 159,676 lines of residential property sales from 2020 to 2023. It included 18 columns providing details such as property descriptions, address, location (by district), size of property (in square foot), unit cost per square foot of the sales price, floor of the flat, and sale prices on specific dates. The dataset was found on Kaggle.[[10]](#footnote-10) In the Python coding this dataset is referred to as the "dataset."

### Macroeconomic Factors Dataset

### This dataset contained 15 macroeconomic factors (described below), which according to research could impact on property prices and property price changes. These macroeconomic factors were established based on the research conducted via the literature review. The data was collected using publicly available quarterly information from the Census and Statistics Department of the Hong Kong Government[[11]](#footnote-11) for the same period of 2020 to 2023. In the Python coding this is referred to as the “property dataset.”

The rationale for selecting these datasets was to combine the detailed transactional data (as the “dataset”) and the high-level macroeconomic dataset (as the “property dataset”) to determine the fundamental and macroeconomic factors influencing property prices and price changes.

The key macroeconomic data was established through a literature review, as previously mentioned.

In the Python code, the merged dataset is called the "merged\_dataset."

## Review of the merged\_dataset

Before conducting machine learning, it was crucial to thoroughly understand the `merged\_dataset` and the types of data it contained to determine its relevance for predicting property prices and property price changes. This analysis revealed the following issues:

### Data Size and Structure: The dataset consisted of 159,676 rows of property transactions across 31 columns.

### Data Types: The data included various types, such as integers, objects, and booleans. However, some columns that should have been floats or integers were incorrectly categorized as objects.

### Data Type Anomalies: For some unknown reason, the merging of the databases caused the float data in the macroeconomic section to be converted into objects.

### Data Type Mismatches: Certain columns were incorrectly assigned as objects when they should have been integers or floats, such as the column representing the size of the flats.

### Price Data Issues: The price data appeared to be rounded and was not calculated as a function of the area multiplied by the unit rate per square foot.

### Descriptive Data: Some columns contained detailed descriptions of the properties in some detail and others had no data filled in, this data was incomplete and descriptive, and it was considered this data, as a whole, would not be directly useful for the analysis.

### Null Values: A significant number of null values were found in columns related to towers, flats, phases, and blocks. This reflects that not all properties were in towers, phases, or blocks, and some were not described as flats (possibly houses).

### Public and Rental Housing Data: The dataset included 1,554 rows of public housing data and 48,267 rows of rental data, which would need to be handled differently from private property sales data.

### Property Size Variation: The size of the properties varied significantly, from less than 500 square feet to over 1,000 square feet, which is considered large by Hong Kong standards.

### Missing Data: Some data was missing for certain districts, particularly during the period of 2020 and 2021.

## Identification of targets and factors

The objective was to identify the factors that could influence property prices and property price changes of residential property and the merged\_dataset contained three ways this objective could be established. These were:

### For each property transaction there was the sale price of the property on a specific day (the “price”);

### For each property transaction there was a unit rate of the property (in Hong Kong $ per square foot) (the unit\_rate); and

### For each property transaction there was the year-on-year price change (called “PC”) based on the quarter the property was sold.

The research suggested that the most significant factors affecting property prices were the size of the property (in square feet) and its location.

These two factors were expected to be the most fundamental factors in determining residential property prices and by extension their cost per square foot. However, the price change of the property could be a result of other macroeconomic factors.

For this reason, it was decided that in analyzing residential property prices and price changes the following targets would be used in the machine leaning algorithms as follows:

1. The property price at a specific time (this was called the “new\_price” for the reason set out below);
2. The unit rate of the property in Hong Kong dollars per square foot (which was called unit\_rate); and
3. The price change (called “PC”) of the property.

Each target would be assessed by the machine learning algorithms separately.

## Data Examination

Before conducting machine learning, it was crucial to thoroughly understand the `merged\_dataset` and the types of data it contained to determine its relevance for predicting property prices and property price changes. This analysis revealed the following issues:

### Data Size and Structure: The dataset consisted of 159,676 rows of property transactions across 31 columns.

### Data Types: The data included various types, such as integers, objects, and booleans. However, some columns that should have been floats or integers were incorrectly categorized as objects.

### Data Type Anomalies: For some unknown reason, the merging of the databases caused the float data in the macroeconomic section to be converted into objects.

### Data Type Mismatches: Certain columns were incorrectly assigned as objects when they should have been integers or floats, such as the column representing the size of the flats.

### Price Data Issues: The price data appeared to be rounded and was not calculated as a function of the area multiplied by the unit rate per square foot.

### Descriptive Data: Some columns contained detailed descriptions of the properties in some detail and others had no data filled in, this data was incomplete and descriptive, and it was considered this data, as a whole, would not be directly useful for the analysis.

### Null Values: A significant number of null values were found in columns related to towers, flats, phases, and blocks. This reflects that not all properties were in towers, phases, or blocks, and some were not described as flats (possibly houses).

### Public and Rental Housing Data: The dataset included 1,554 rows of public housing data and 48,267 rows of rental data, which would need to be handled differently from private property sales data.

### Property Size Variation: The size of the properties varied significantly, from less than 500 square feet to over 1,000 square feet, which is considered large by Hong Kong standards.

### Missing Data: Some data was missing for certain districts, particularly during the period of 2020 and 2021.

## Data Cleaning

As set out above, several issues were identified with the merged\_dataset which was planned to be used for the machine learning analysis. This resulted a requirement to clean the data before analysis.

While on first look the property transactions appeared to be comprehensive and record all property sales over a period of three years; in reality it was incomplete and inconsistent for all districts. In particular, for the years 2020 and 2021, it appears that not all property transactions seemed to have been inputted or recorded consistently.

 The property transactions dataset also included irrelevant information that did not contribute to understanding prices, such as tower, block, and phase. Not all properties in Hong Kong have separate towers, blocks, and phases, and this descriptive data was sometimes left blank. Given this inconsistency, these fields needed to be removed from the dataset.

Other information which may affect property prices such as sea view, distance to schools, distance to public transport (the underground) were not included in the dataset.

Additionally, to enable an analysis using machine learning, for those columns which would be used, the numerical data types needed to be converted to integer or float formats.

The macroeconomic factors dataset posed challenges as well. This information was not freely available as a single dataset and had to be compiled from the Hong Kong government census and statistics website.[[12]](#footnote-12) The government provided macroeconomic data only quarterly. Given that property transactions occurred daily, the 159,676 transactions needed to be aligned with the corresponding quarters to align the macroeconomic factors data with each sales transaction.

## Data Cleaning Techniques Used

To create a merged dataset suitable for analysis using machine learning, the data was pivoted, reformatted, analyzed, cleaned and managed using the following techniques:

### Alignment of the two datasets: It was necessary to ensure a proper alignment of the two datasets by reorienting (or pivoting) the columns in the macroeconomic factors dataset and merging them so that property transaction dates would be aligned with the relevant quarters of the macroeconomic data.

### Removal of Irrelevant Information: The property transactions dataset included information not pertinent to property prices, such as rental prices. Since the project concerns property sales, rental prices were irrelevant, so these entries were removed from the database.

### Removal of Public Housing Data: The property transactions dataset included sales of public housing. The project was to analyze private residential housing in Hong Kong. Public housing is subsidized by the government, so it does not reflect true market prices. Therefore, the public housing entries would skewer the results, so this public housing data was removed from the dataset. Data Type Conversion: For relevant columns, it was necessary to change the data types from objects to integers or floats to ensure compatibility with various machine learning algorithms. This was particularly the case in the saleable area, which was an object data type, and all of the macroeconomic indicators were changed from floating data type to an object data type during the merging process of the two datasets.

### Handling Missing Values: It was necessary to address missing (or null) values from the property transactions dataset. From analysis of the null values, these were in the columns of “tower,” “flat,” “phase,” and “block.” When reviewing  these descriptions, they appeared to be inconsistent and irrelevant to all of the properties, and these columns were deleted, eradicating the problem of null values.

### The Problem with the Prices. The price column in the property transactions dataset needed to be corrected. This was because the property price values were rounded figures as they did not reflect the area multiplied by the unit rate per square foot. For this reason, the actual prices were re-calculated using area multiplied by the unit rate to ensure accuracy. In the merged dataset, a new column (with the corrected price) was called “new\_price” was created, and the old column called “price” was deleted or dropped.

### Addressing outliers. Part of the data analysis was to address outliers in the data, as some machine learning algorithms are sensitive to outliers and can cause errors in prediction. In this project, the Interquartile Range (IQR) of +1.5 was used to detect outliers.[[13]](#footnote-13)

### Normalization of Data: The values in the merged dataset were normalized as some machine learning algorithms are sensitive to scale and could skew the results. In this project, a MinMaxScaler[[14]](#footnote-14) was used to scale the data between 0 and 1.

## Feature Engineering

Based on the data in the merged\_database it was also possible to create new features which could then be used in the analysis. The new features were:

### One Hot Encoding for the Districts. The research suggested that the most important predictors of property prices is the area (size of the property) and  district (or location) of the property. For this project, this was considered an essential factor in any machine learning analysis. Therefore, as there were 18 districts in the dataset, it was decided that each district would be one hot encoded so it could be used later in the machine learning algorithms.

### Creating regions from districts. The merged\_dataset contained 18 districts and it was felt that it would be useful to group these districts into 4 regions which are generally used in Hong Kong. These regions were Hong Kong Island District (HK), Kowloon (KLN), New Territories East (NTEast) and New Territories West (NTWest) for further analysis.

### Classifying flat size. The merged\_dataset contained different sizes of property. Property in Hong Kong is small by international standards and it was felt it would be useful to group properties by size such as: “small” (less than 500 square foot), “medium” (500 to 1000 square foot), “large” (1000 to 1500 square foot) and “very large” (greater than 1500 square foot).

These additional features were created to enrich the dataset and improve the predictive power of the machine learning models.

### 

# Data Analysis Results

## Preliminary data analysis

The primary objective of this project was to analyze the factors influencing property prices and their fluctuations using machine learning techniques. However, before diving into the complex modeling and prediction tasks, it was essential to conduct a thorough preliminary data analysis of the `merged\_dataset`. This initial step was crucial for understanding the data's structure, identifying any anomalies, and ensuring the quality of the dataset before applying machine learning algorithms.

To achieve this, various exploratory data techniques were employed using the Python libraries such as Pandas and Matplotlib.

This enable an in-depth examination of the dataset, including the distribution of key variables, trends over time, and potential correlations between features. The results of this analysis were visualized graphically, providing clear insights into the underlying patterns within the data.

These enabled the identification of problem areas to guide the subsequent machine learning modeling process.

The following data analysis was then done as follows: [GRAPHS IN PYTHON CODING]

### Review of the price per square foot of the movement of private property sold by Quarter

[GRAPH]

The graph illustrates that property prices initially dipped in mid-2020, reflecting the impact of COVID-19,[[15]](#footnote-15) but recovered by late 2020, peaking in mid-2021 as global economies reopened. However, the property market did not sustain this recovery, possibly due to economic challenges in Hong Kong and China (affecting sentiment), diminished autonomy of Hong Kong, and ongoing USA-China trade tensions.[[16]](#footnote-16) The slight recovery in early 2023 suggests some market stabilization, but it indicates only a modest improvement, with prices still not reaching 2020 levels.

### Review of the number of transactions of sold properties per quarter

### [GRAPH]

### The shows a gradual increase in property transactions from a low base in 2020 to 2023.

### Review of the the distribution of unit rate for private properties

[GRAPH]

The graph shows a right-skewed distribution of unit rates for private properties, with most transactions clustered below the mean price of approximately HK$16,423.51 per square foot. This suggests that buyers are more interested in buying low and mid range properties and a reflection of the overall high cost of property in Hong Kong.

### Review of the mean unit rate of sold properties per quarter

[GRAPH]

The graph shows the mean unit rate of sold properties fluctuates slightly over time, with peaks around Q4 2020 and Q2 202, followed by a gradual decline in more recent quarters. The overall mean unit rate across all quarters is approximately HK$16,899.48. This suggests some variability in property prices over time, with a general downward trend towards the end of the period shown.

### Review of the number of transactions of sold properties by size and quarter

[GRAPH]

The graph shows an increase in the total number of transactions has increased with medium and large properties driving much of the increase.

### Reviewing a box plot of unit rate by property size of sold properties

[GRAPH]

The box plot shows that unit rates generally increase with property size, with larger properties displaying greater variability and more outliers. "Very Large" properties have the highest median rates and widest spread, indicating diverse pricing probably influenced by other factors such as prestige, specification, design, sea view and other features.

### Review of the price per square foot and the movement of sold property by sizes and by quarter

[GRAPH]

This shows that "Very Large" properties consistently have the highest price per square foot, with noticeable fluctuations. "Small" and "Medium" properties have more stable and lower price trends, with less variability over time. This indicates that larger properties experience more price volatility, likely due to their luxury nature and sensitivity to market conditions.

### Review of the distribution of sold property unit rates by district and regions

[GRAPH]

This graph illustrates significant variability in pricing by district. Central and Western Districts have the highest median rates and widest ranges, while districts like Tuen Mun and Yuen Long are more affordable.

The box plot of the regions differences in property unit rates, show that  Hong Kong Island has the highest and most variable prices, followed by Kowloon. The New Territories regions exhibit lower, more consistent unit rates with fewer outliers.

### Reviewing the number of sales transactions of sold properties by district and Quarter

[GRAPH]

The chart indicates that property sales transactions vary significantly across districts, with some districts consistently contributing more to the total. It highlights differences in market activity between these districts.

## Checking the completeness of the data

It was necessary to review the completeness of the dataset and determine if there was any missing data in the `merged\_dataset`. The methodology involved analyzing how the data was distributed over the years by examining the average price data for each quarter. This analysis identified that many districts had missing data. I prepared a heatmap of the results as follows:

[GRAPH]

The heatmap shows that the data is more incomplete in the earlier years (2020 and 2021), as indicated by the extensive blue areas. In contrast, the data becomes more complete in 2022 and 2023, as reflected by the predominance of red. This suggests an overall improvement in data quality during these later years.

The lack of data in earlier periods could be attributed to disruptions caused by the COVID-19 pandemic, which likely affected data collection processes. As conditions stabilized, there was a stronger emphasis on data collection and accuracy, resulting in more complete and reliable data in 2022 and 2023.

It is difficult to reconcile the notion of zero property sales reflected in the data during the COVID-19 period in some districts, especially since other districts still recorded sales. This suggests that some agent offices in certain districts likely did not complete the database in the early and or COVID-19 period.

Due to the missing data in the database for all districts, especially in the earlier years it would likely skew the analysis and results and will pose challenges for the statistical analysis.

The gaps in data may compromise the validity of any trends or patterns identified, as the analysis might not accurately reflect the true state of the market across all districts and regions. This could severely hinder the ability to compare districts and regions effectively, as some areas may be overrepresented while others are underrepresented due to missing data. As a result, the conclusions drawn from such analysis could be misleading or incorrect, potentially leading to flawed decisions based on inaccurate insights.

In the application of machine learning, these data gaps further complicate the situation. The quality and completeness of data are critical for training robust machine learning models. Missing data can lead to reduced model accuracy, as the algorithms may struggle to identify meaningful patterns when essential information is absent. This increases the risk of overfitting, where the model performs well on the training data but is less accurate on new data leading to the model not predictions property prices and or price changes accurately.

Additionally, the lack of representative data from certain periods or regions can hinder the model's ability to generalize and could result in unreliable predictions and low accuracy.

The problems of the data based are highlighted by looking at the changes in price by district as some districts do not appear in the analysis as the graph below:

[GRAPH]

## Correlation Analysis

As the objective was to consider the factors which influenced the property price or price changes then understanding the correlation of these target(s) with the other variables would be important.

A correlation analysis using a heat map was performed which visualized correlations between the three specific targets[[17]](#footnote-17)and the macroeconomic factors. The results were as follows:

* + 1. The unit rate of the property (“unit\_rate”) which is the unit rate or cost per square foot of the purchase price of the property when compared with the macroeconomic factors showed generally weak correlations. This suggests that unit rate is not strongly influenced by any of the macroeconomic factors in the model. This may be because unit rate is influenced by other factors such as size and location and or additional features may be required or influenced by a combination of factors rather than any single variable. [GRAPH]
    2. The price (“new\_price”) which was the sale price of the properties. The heat map reveals generally weak correlations between the macroeconomic variables which imply that the new price cannot  be predicted using these variables alone for the same reasons as the unit rate above.
    3. The Price Change (“PC”) of the properties. The heat map reveals strong positive correlations with “IR” (interest rates), “MW” (monthly wages), and “SM” (stock market) which suggests that these are key predictors for property price changes. There are also strong negative correlations with “CPI” (consumer price index) and “HS” (housing starts) which indicate that higher consumer prices and lower housing starts are associated with lower property price increases. These insights imply that many of the macroeconomic factors in the model are good indicators of price changes. [GRAPH]

The weak correlations of unit rate and new price with macroeconomic indices are contrasted with the strong correlations of the price change and suggest that unit rate and new price are influenced by other factors not captured by these macroeconomic indicators. This may be because unit rate and new price` are influenced by other key factors such as property size and location (district).

## Multicollinearity Analysis

Within the merged dataset, there was a concern about a potential crossover between unit rate, saleable area, and price. Specifically, if you have two of these variables, you can derive the third, which can cause problems in determining which factors are actually correlated.

Addressing multicollinearity would be necessary to improve model stability and prediction accuracy. This involves using techniques like Ridge and Lasso regressions, which add regularization to reduce overfitting and enhance interpretability.

The results of the multicollinearity analysis are as follows:

### Unit Rate: High VIF values indicate significant multicollinearity. Both Ridge and Lasso regression models perform poorly, with high MSE (~7,422,559) and low R² (~0.07), suggesting they fail to effectively predict the unit rate.

### New Price: The analysis shows significant multicollinearity. Ridge and Lasso models also perform poorly, with extremely high MSE (~1,734,676,577) and low R² (~0.09), indicating poor prediction capability for the price.

### Price Change: Despite high VIF values, Ridge Regression performs excellently with very low MSE (~0.0716) and high R² (~0.999). Lasso Regression is also good but slightly less effective, with MSE (~0.557) and R² (~0.99).

Based on the above analysis, the analysis of the property database using Lasso and Ridge regression models it shows for `unit\_rate` and `new\_price`, both models exhibit low R² values, indicating limited suggesting that feature in the dates base don to capture the changes in Unit\_rate and Price and that these targets are driven bu other feature (such and size location).

However, the model performs well for the `PC` (Price Change), with high R² values indicating a good fit (and potential overfitting). This suggests that the features used in the database are well-suited to predicting price changes of the properties. Overall, while the models are only effective for PI and furtherenhancements are needed for the other targets to achieve more reliable predictions.

# Machine Learning Results

Following the data cleaning and analysis, the following machine learning algorithms we used to review the data. Each machine learning algorithm was assessed based on the three targets.

## Regression Analysis[[18]](#footnote-18)

## Regression analysis was selected as it is well-suited for analyzing and predicting property prices as it has the ability to handle the multiple property variables simultaneously, identifying and quantifying the relationships between property prices and various macroeconomic factors.

The regression analyis results were for each target are as follows:

### Unit rate. The regression analysis explains that 65.8% of the variance for unit rate is accounted for by the model (R-squared = 0.658), with key predictors being `saleable\_area (cost per square foot), the size of the flat and `floor` (height), among others. However, the model shows signs of multicollinearity (as discussed above), and exploring additional predictors could enhance model accuracy and reliability. The scatter plot shows a general trend along the line of equality, indicating that the model's predictions are generally in the right range, but there is considerable spread, suggesting that the predictions are not perfect and there is room for improvement.

### New Price. The regression analysis explains that 66.8% of the variance in the new price is accounted for by the model (R-squared = 0.667), with key predictors being `saleable\_area (cost per square foot), among others. The residuals vs fitted plot shows that the errors are spread unevenly, indicating that the model's predictions are not consistently accurate. The QQ plot shows that the errors do not follow a normal pattern. These issues mean that the model's assumptions are not fully met, which can affect the reliability of its predictions. There are issues of multicollinearity which need to be addressed to improve model accuracy and reliability.

### Price Change (PC). The regression analysis explains that 99.8% of the variance in price changes are accounted for by factors in the model, indicating an excellent fit. Most factors are significant. The Actual vs Predicted PI plot shows that the model's predictions closely match the actual values. However, the Residuals vs Fitted plot reveals some outliers, and the QQ plot indicates that the errors are not perfectly normal. Again, there are signs of multicollinearity, however, the model is highly accurate, with minimal underlying problems.

The regression results show, as expected, the property prices are most influenced by area and location, followed by floor and the macroeconomic factors are important for addressing price changes.

The regression model shows that macroeconomic factors crucially explain 99.8% of `price change variance, reinforcing their importance in price changes.

The models could be improved by addressing multicollinearity and by considering other relevant factors that could explain the remaining variance in property prices and price changes.

## Decision Trees Analysis

Decision Trees was used for this project as it can be used to capture interactions between various macroeconomic factors and the impact on property values. By splitting the data into branches based on feature values to provide clear results.

The results of the Decision Tree analysis for the three targets are as follows:

### Unit Rate. The `unit\_rate` model shows moderate performance with an MSE[[19]](#footnote-19) and a R2 score.[[20]](#footnote-20) The scatter plot indicates a dispersed pattern of predicted vs. actual values, suggesting that the macroeconomic factors are not very useful in predicting the unit rate. Feature engineering might be necessary to improve the predictions.

### New Price. For `new\_price`, the model's performance is also shows a moderate MSE and R2 score[[21]](#footnote-21) with the scatter plot showing significant dispersion, indicating the current features are not effective in accurately predicting the new price. Considering ensemble methods like Random Forests could enhance the model's accuracy.

### Price Changes. The model demonstrates excellent performance with an MSE and a R2 score of 1.0,[[22]](#footnote-22) with the graph showing a near-perfect alignment. This suggests that the features accurately capture the relationship between factors and price changes, resulting in highly accurate predictions. This model is well-suited for this target variable.

To improve the model's accuracy, it may be necessary to incorporate other factors. Important factors to consider include the square footage of the property, as larger properties often have higher unit rates and prices, and the location (district). Other potentially important factors not included in the dataset could be the age of the property, its condition, proximity to transport links (particularly the MTR in Hong Kong), school quality, and crime rate.

## XGBoost Analysis[[23]](#footnote-23)

XGBoost was chosen as a machine learning algorithm due to its high predictive accuracy and ability to capture complex non-linear relationships. Its built-in regularization minimizes overfitting, ensuring robust generalization, and it allows for effective hyperparameter tuning. Two analyses were conducted: one on an untuned model, followed by hyperparameter tuning and a re-run. The results are as follows:

The untuned model is as follows:

### The XGBoost learning curves for `unit\_rate` and `new\_price` show that as the number of training examples increases, the training error decreases and stabilizes, indicating effective learning. The validation error starts high but converges with the training error, suggesting that the models are generalizing well without significant overfitting. However, the complexity of the data might require additional features or further parameter tuning to enhance prediction accuracy.

### For the `PI` (Price Index), the learning curve reveals excellent model performance. The training error is very low, and the validation error quickly drops to align closely with it, indicating that the model fits the data well and captures the underlying patterns effectively. This suggests that the features used are well-suited for predicting the PI, with a low risk of overfitting. Overall, while the `PI` model performs exceptionally, the `unit\_rate` and `new\_price` models could benefit from further refinement.

To enhance prediction accuracy it was decided to optimize the model parameters for `unit\_rate` and `new\_price` and PI` by hyperparameter optimization using GridSearchCV. The hyperparameter tuning analysis provided the following for each target:

### For all three targets of price, unit rate and price changes all showed the best results using 3-fold cross-validation with a total of 729 fits.

### For price the best tuning was: `max\_depth=9`, `eta=0.1`, and `subsample=1.0`, yielding an R-squared of 0.984;

### For unit rate` the best tuning was `max\_depth=9`, `eta=0.2`, with an R-squared of 0.973. The low mean squared error (MSE) values indicate high accuracy, particularly for `PI`. Overall, the models demonstrate strong predictive power and effective hyperparameter optimization; and

### For Price Changes the best tuning was `max\_depth=6`, `eta=0.2`, and `colsample\_bytree=0.6`, achieving an almost perfect R-squared of 1.0;

Based on the hyperparamter results the XGBoost analysis was re-run and the comparison is presented in a table format:

Table 1 – Comparison of XGBoost Results using tuned and untuned hyper parameter models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Metric** | **Unit Rate** | | **Price** | | **Price Change** | |
|  | **Tuned** | **Untuned** | **Tuned** | **Untuned** | **Tuned** | **Untuned** |
| Training Error | Low | High | Low | Moderate | Low | Moderate |
| Validation Error | Higher | High | Low | High | Low | High |
| Generalisation | Good | Poor | Excellent | Poor | Excellent | Moderate |
| Overfitting | Some indication | High | Minimal | Present | Minimal | Present |
| Model performance | High | Low | High | Medium | Very High | Moderate |

The comparison of tuned and untuned XGBoost models reveals that tuned models consistently exhibit lower training and validation errors, indicating better generalization and minimal overfitting. Overall, tuned models demonstrate significantly improved predictive performance across all dependent variables, while untuned models struggle with accuracy and generalization.

## MLP Analysis[[24]](#footnote-24)

In this Project it was hoped that MPL would be a good choice for forecasting property prices, price changes and the unit rate of properties as the approach can capture non-linear relationships in the merged dataset. The MPL analysis is as follows:

### Unit rate. The learning curve shows a high initial error that decreases sharply as more training examples are added. The rapid convergence of training and validation errors suggests that the MLP model is learning effectively and generalizes well with the available data. This indicates that the model complexity is appropriate, and the features are relevant for predicting `unit\_rate`.

### Price. The learning curve starts with a high initial error, and both training and validation errors plateau over a range of examples. A sudden drop in validation error at the end may indicate potential overfitting or a significant improvement. Further investigation, potentially with cross-validation, is needed to ensure the model's reliability and to address any anomalies.

### Price Change. The learning curve shows a high initial error that decreases rapidly, with both errors stabilizing at low levels. This suggests the MLP model is well-fitted and generalizes effectively. The consistent low error indicates that the model's complexity aligns well with the dataset, and it captures the underlying patterns necessary for predicting `PI` without overfitting.

# Discussion on results

* 1. **Summary of Results**

### Traditional Data Analysis

XXX

### Machine Learning

XXX

* 1. **Comparison of Machine Learning Results**

From the analysis of various machine learning methodologies, the results were as follows:

Table 2 - Comparison of Machine Learning Results

|  |  |  |  |
| --- | --- | --- | --- |
|  | Regression Analysis | | |
| Target | Unit Rate | Price | Price Change |
| Variance | 65.8% | 66.8% | 99.8% |
| R-squared | 0.658 | 0.667 | 0.667 |
| Key predictors | Saleable  area / size/ floor | Saleable  area | Most factors are significant |
| Scatter Plot | A shows a general trend along the line of equality, indicating that the model's predictions are generally in the right range, but there is considerable spread | Shows that the errors are spread unevenly, indicating and that the errors do not follow a normal pattern | The model's predictions closely match the actual values and the Residuals vs Fitted plot reveals some outliers, and that the errors are not perfectly normal |
| Problems | Signs of multicollinearity | Signs of multicollinearity | Some outliers |
| Indications | The predictions are not perfect and there is room for improvement | The predictions are not consistently accurate | Some errors and there are signs of multicollinearity, |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Decision Tree Analysis | | |
| Target | Unit Rate | Price | Price Change |
| MSE | 41,282,658.68 | 108,670,131 | 100% |
| R-squared | 0.4827 | 0.43 | 1.00 |
| Are macroeconomic indicators useful | The macroeconomic factors are not very useful in predicting the unit rate | The current features are not effective in accurately predicting the new price | the features accurately capture the relationship between factors and price changes |
| Percentage of factors that are useful | 51.73% | 43.08% | 100% |
| Scatter Plot | A dispersed pattern of predicted vs. actual values, suggesting that the macroeconomic factors are not very useful | A showing significant dispersion, indicating the current features are not effective in accurately predicting the new price | Showing a near-perfect alignment |
| Problems | NA | Signs of multicollinearity | Some outliers |
| Indications | Suggest feature engineering | Suggest feature engineering | Highly accurate |

|  |  |  |  |
| --- | --- | --- | --- |
|  | XGBoost Analysis (untuned) | | |
| Target | Unit Rate | Price | Price Change |
| Curve | As the number of training examples increases, the training error decreases and stabilizes | As the number of training examples increases, the training error decreases and stabilizes | The training error is very low, and the validation error quickly drops to align closely with it, indicating that the model fits the data well and captures the underlying patterns effectively |
| Model | Indicates effective learning | Indicates effective learning | Excellent performance and the features used are well-suited for predicting the PI, with a low risk of overfitting |
| Signs of overfitting | No | No | Most factors are significant |
| Factors relevant | Not relevant | Not relevant |  |
| Problems | The complexity of the data might require additional features or further parameter tuning to enhance prediction accuracy | The complexity of the data might require additional features or further parameter tuning to enhance prediction accuracy | Could benefit from refinement |

|  |  |  |  |
| --- | --- | --- | --- |
|  | MPL Analysis | | |
| Target | Unit Rate | Price | Price Change |
| Learning Curve | Shows a high initial error that decreases sharply as more training examples are added there is a rapid convergence of training and validation errors | Starts with a high initial error, and both training and validation errors plateau over a range of examples. | Starts with a high initial error, and both training and validation errors plateau over a range of examples. |
| Model | Model is learning effectively and generalizes well with the available data | Generalizing | The model is well-fitted and generalizes effectively |
| Features | The features are relevant |  | The consistent low error indicates that the model's complexity aligns well with the dataset |
| Problems | Model is appropriate | A sudden drop in validation error at the end may indicate potential overfitting or a significant improvement and potentially with cross-validation, is needed to ensure the model's reliability and to address any anomalies | A sudden drop in validation error at the end may indicate potential overfitting or a significant improvement |

In summary, regression shows moderate accuracy but struggles with multicollinearity and inconsistent predictions. Decision Trees have lower R-squared for Unit Rate and Price but excel in predicting Price changes, suggesting feature engineering. XGBoost demonstrates effective learning and outstanding performance for Price Changes, with minimal overfitting risk, though it may require tuning. MPL models generalize well, aligning with data complexity, but potential overfitting is indicated by sudden error drops. Overall, XGBoost and MPL are particularly strong for Price Change predictions, while Decision Trees effectively capture this relationship when features are well-aligned.

## Problems and Challenges

One of the primary challenges was obtaining a complete dataset of property transactions in Hong Kong.

The selected dataset covered only a three-year period, specifically during the COVID-19 pandemic, when transaction volumes were significantly low and not representative of typical market conditions. During this time, there was a notable reduction and fluctuation in property prices across various districts.

The incompleteness of the data, particularly for certain districts, hindered comparisons. This issue was especially pronounced in the early months of 2020 and 2021, likely due to the company's implementation of transaction recording protocols and reduced staff involvement during the pandemic.

Given sufficient time and resources, it would be possible to compile a comprehensive record of all property transactions, as these are documented in the land registry. However, this would require a more extensive dataset, ideally spanning 10 to 20 years.

While other datasets were available, they also presented challenges, such as incomplete coverage and inconsistencies across different areas of Hong Kong. For example other dataset reviewed[[25]](#footnote-25)also had their own limitations, further complicating the analysis.

# Conclusion

## Model Reliability

As a result of the constraints related to the data, the insights gained from the machine learning analysis were inconclusive and, in some respect resulted in limited usefulness.

## Future Research and last thoughts

This exercise demonstrated a real-world example of the challenges associated with datasets, including the manipulation and cleaning of data, as well as the issues surrounding the completeness of datasets over extended periods needed for trend analysis.

In the future, it would be possible to develop a significantly larger dataset of property transactions spanning many years. However, this would likely involve manually inputting hundreds of thousands of transactions, making it extremely labor-intensive. Given the extensive effort required, it remains questionable whether this would be a worthwhile endeavor.

Reference list / Bibliography

1. “7 Emerging Trends in Hong Kong’s Real Estate Market: 2024 Outlook” (2024), 25 May. Available at: http://www.senatormensch.com/7-emerging-trends-in-hong-kongs-real-estate-market-2024-outlook/ (Accessed: June 25, 2024).
2. “An Overview of Loan-to-value Ratio for Mortgages in Hong Kong” (2024), 28 February. Available at: https://www.spacious.hk/en/blog/an-overview-of-loan-to-value-ratio-for-mortgages-in-hong-kong (Accessed: June 26, 2024).
3. *Bloomberg.Com* (2023) “Hong Kong Home Prices to Slump Through 2024, Morgan Stanley Says,” 12 October. Available at: https://www.bloomberg.com/news/articles/2023-10-12/morgan-stanley-now-sees-hong-kong-home-price-slump-through-2024 (Accessed: June 25, 2024).
4. *Bloomberg.Com* (2024) “Hong Kong’s Property Market Faces Headwinds Even After Curbs Are Scrapped,” 28 February. Available at: https://www.bloomberg.com/news/articles/2024-02-28/hong-kong-s-push-to-lift-property-market-faces-uphill-battle (Accessed: June 25, 2024).
5. *Five factors that may sway Hong Kong home prices in 2022* (2021). Available at: https://www.scmp.com/business/article/3161224/five-biggest-factors-may-influence-hong-kong-home-prices-year-ahead (Accessed: June 25, 2024).
6. Frank, K. (no date) *Chinese mainland and Hong Kong property market 2024 forecasts - Knight Frank*. Available at: https://www.knightfrank.com.hk/blog/2024/06/12/chinese-mainland-and-hong-kong-property-market-2024-forecasts (Accessed: June 25, 2024).
7. Gerlach, S. and Peng, W. (2005) “Bank lending and property prices in Hong Kong,” *Journal of Banking & Finance*, 29(2), pp. 461–481. Available at: https://doi.org/10.1016/j.jbankfin.2004.05.015.
8. Guide, G.P. (2024a) *Hong Kong’s housing market conditions continue to deteriorate (Q2, 2024)*. Available at: https://www.globalpropertyguide.com/asia/hong-kong (Accessed: June 26, 2024).
9. Guide, G.P. (2024b) *Hong Kong’s housing market conditions continue to deteriorate (Q2, 2024)*. Available at: https://www.globalpropertyguide.com/asia/hong-kong (Accessed: July 31, 2024).
10. Herring, E. (2018) *Average property price by country around the world*. Available at: https://www.finder.com/uk/mortgages/world-cost-of-a-flat (Accessed: August 4, 2024).
11. *Hong Kong Budget 2024-25: Key Highlights of Property Measures | Insights | Mayer Brown* (no date). Available at: https://www.mayerbrown.com/en/insights/publications/2024/02/hong-kong-budget-2024-25-key-highlights-of-property-measures (Accessed: June 25, 2024).
12. *Hong Kong: Census and Statistics Department : Housing and Property 2024* (no date). Available at: https://www.censtatd.gov.hk/en/scode100.html (Accessed: June 25, 2024).
13. *Hong Kong homes remain most unaffordable in world for 14th year: research* (2024). Available at: https://www.scmp.com/business/article/3266634/hong-kong-homes-remain-most-unaffordable-world-14th-year-research (Accessed: June 25, 2024).
14. *Hong Kong Major Report - Hong Kong Market Outlook 2024* (no date). Available at: https://www.cbre.com.hk/insights/reports/hong-kong-major-report-hong-kong-market-outlook-2024 (Accessed: June 25, 2024).
15. *Hong Kong Real Estate as an Investment* (no date). Available at: https://www.streetdirectory.com/travel\_guide/67409/asia\_properties/hong\_kong\_real\_estate\_as\_an\_investment.html (Accessed: July 31, 2024).
16. “Hong Kong’s property markets in 2024 | Savills Prospect” (2023), 6 December. Available at: https://www.savills.com/prospects/hong-kong-property-markets-in-2024.html (Accessed: June 25, 2024).
17. *Hong Kong’s Residential Property Market Faces More Challenges in 2024* (no date). Available at: https://www.worldpropertyjournal.com/real-estate-news/hong-kong/hong-kong-real-estate-news-2024-hong-kong-property-market-predictions-jll-hong-kong-property-market-forecast-for-2024-hong-kong-home-sales-predictions-13978.php (Accessed: June 25, 2024).
18. Kong, D.H. (2024) “Hong Kong Property Review 2024 unveils detailed market insights,” 26 April. Available at: https://www.dimsumdaily.hk/hong-kong-property-review-2024-unveils-detailed-market-insights-and-future-projections/ (Accessed: June 25, 2024).
19. Li, J. *et al.* (2022) “Assessing economic, social and environmental impacts on housing prices in Hong Kong: a time-series study of 2006, 2011 and 2016,” *Journal of Housing and the Built Environment*, 37(3), pp. 1433–1457. Available at: https://doi.org/10.1007/s10901-021-09898-x.
20. *Rating and Valuation Department - Publications* (no date). Available at: https://www.rvd.gov.hk/en/publications/property\_market\_statistics.html (Accessed: June 25, 2024).
21. “Savills Research: Property Markets in 2024” (no date). Available at: https://www.savills.com/prospects/tag/property-markets-in-2024/ (Accessed: June 25, 2024).
22. Standard, T. (no date) *New round of subsidized public housing prices at HK$750,000 to HK$2.71 million &nbsp;* Available at: https://www.thestandard.com.hk/breaking-news/section/4/201718/New-round-of-subsidized-public-housing-prices-at-HK$750,000-to-HK$2.71-million-%C2%A0 (Accessed: July 31, 2024).
23. *Table 510-60001 : Consumer Price Indices (October 2019 – September 2020 = 100)* (no date). Available at: https://www.censtatd.gov.hk/en/web\_table.html?id=510-60001# (Accessed: June 26, 2024).
24. *The Evolution of the Hong Kong Property Market* (2024a). Available at: https://insights.grcglobalgroup.com/the-evolution-of-the-hong-kong-property-market/ (Accessed: July 30, 2024).
25. *The Evolution of the Hong Kong Property Market* (2024b). Available at: https://insights.grcglobalgroup.com/the-evolution-of-the-hong-kong-property-market/ (Accessed: July 31, 2024).
26. Wong, T.C., Tsang, A. and Kong, S. (2014) “How Does Loan-to-Value Policy Strengthen Bankss Resilience to Property Price Shocks Evidence from Hong Kong,” *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.2396965.

Appendix 1

1. The following were used: Decision Trees, Regression, XGBoost (Extreme Gradient Boosting), and MLP (Multi-Layer Perception) [↑](#footnote-ref-1)
2. This dataset provides a historical housing prices scraped from Centaline Property Hong Kong which is one of the largest real estate agencies in Hong Kong. The dataset includes information on the date of the transaction, the property address, floor plan, saleable area, unit rate, source, and district. The dataset covers a period of time spanning several years, allowing for analysis of trends and changes in the Hong Kong housing market. See https://www.kaggle.com/datasets/cyrusttf/hong-kong-housing-price-2020-2023 [↑](#footnote-ref-2)
3. Global Property Guide (2024). Hong Kong’s housing market remains depressed. Global Property Guide. [online] doi:https://www.globalpropertyguide.com/asia/hong-kong/price-history [↑](#footnote-ref-3)
4. https://en.wikipedia.org/wiki/Geography\_of\_Hong\_Kong#:~:text=The%20area%20of%20Hong%20Kong,part%20of%20%22Greater%20China%22.&text=Hong%20Kong%20has%20a%20total,southwest%20of%20the%20main%20peninsula. [↑](#footnote-ref-4)
5. https://www.finder.com/uk/mortgages/world-cost-of-a-flat [↑](#footnote-ref-5)
6. XXX [↑](#footnote-ref-6)
7. XXX [↑](#footnote-ref-7)
8. XXX [↑](#footnote-ref-8)
9. XXX [↑](#footnote-ref-9)
10. https://www.kaggle.com/datasets/cyrusttf/hong-kong-housing-price-2020-2023 [↑](#footnote-ref-10)
11. See https://www.censtatd.gov.hk/en/page\_32.html [↑](#footnote-ref-11)
12. XXX [↑](#footnote-ref-12)
13. XXX [↑](#footnote-ref-13)
14. XXX [↑](#footnote-ref-14)
15. XXX [↑](#footnote-ref-15)
16. XXX [↑](#footnote-ref-16)
17. See section XXX above [↑](#footnote-ref-17)
18. XXX [↑](#footnote-ref-18)
19. XXX [↑](#footnote-ref-19)
20. XXX [↑](#footnote-ref-20)
21. XXX [↑](#footnote-ref-21)
22. XXX [↑](#footnote-ref-22)
23. XXX [↑](#footnote-ref-23)
24. XXX [↑](#footnote-ref-24)
25. [↑](#footnote-ref-25)