**MSc AI Project (Online) (771952\_A23\_T3A)**

**Project for Module 5**

Title: Report on using machine learning techniques to identify the fundamental and macroeconomic factors affecting property prices and property price increases in Hong Kong.

Student 202249387[[1]](#footnote-1)

[Faculty of Science and Engineering](https://www.hull.ac.uk/faculties/faculty-of-science-and-engineering)

MSc in Artificial Intelligence

University of Hull Online

England, United Kingdom

[timhallworth@netvigator.com](mailto:timhallworth@netvigator.com) and [T.HALLWORTH-2022@hull.ac.uk](mailto:T.HALLWORTH-2022@hull.ac.uk)

**ABSTRACT**

This project used machine learning algorithms[[2]](#footnote-2) to predict residential property prices and price increases in Hong Kong.

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

The project used two databases. One was a large residential sales data set 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. A second data set 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 analyzed 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 increases, but the datasets were incomplete, and improving the data sets would have been time-consuming and beyond the scope of the study.

1. **INTRODUCTION**

From 2008 to 2013, residential property prices skyrocketed by 134%. Despite a market slowdown in the first half of 2014, house prices 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.

Potential land 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 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, and does Hong Kong property still offer opportunities for investors?

By utilizing AI, the project hoped to provide insight into this question.

1. **LITERATURE REVIEW AND METHODOLOGY**

In preparation for this project, a literature review was conducted to investigate the factors affecting residential property prices in Hong Kong. This review included research on academic papers and industry news sources to provide comprehensive insights.

The project utilized two datasets: a large residential data set of sales dataset from a major Hong Kong property agency and a macroeconomic data set 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, decision trees, XGBoost, and multilayer perceptron (MLP), to the merged dataset. Hyperparameter optimization was also performed to enhance prediction accuracy.

1. **FACTORS AFFECTING PROPERTY PRICES**

Like all property worldwide, residential property prices are influenced by two fundamental factors: property location and the size of the property.

However, property price changes are influenced by various macroeconomic factors, which research has identified. This research 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:

3.1 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.

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

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

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

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

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

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

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

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

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

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

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

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

3.14 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.

3.15 PI (Property Price Increase): The historical year-on-year property price increases can provide insights into future price trends.

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

1. **DESCRIPTION OF THE DATASETS USED AND RATIONALE FOR USING THE DATASETS**

The project utilized two datasets and merged them into a single dataset called the merged\_dataset for analysis. The two datasets[[6]](#footnote-6) are as follows:

4.1 Property Transactions Dataset:[[7]](#footnote-7) This dataset contain159,676 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 data set was found on GitHub.[[8]](#footnote-8) There are no ethical and confidentiality issues as this information was published on GitHub and is publically available. In the Python coding this data set is referred to as the "dataset."

4.2 Macroeconomic Factors Dataset:[[9]](#footnote-9) This dataset examines 15 macroeconomic factors (described above), which according to research could impact on property prices and price increases. These macroeconomic factors were established based on the research conducted via the literature review. The data was collected using publicly available quarterly statistics from the Hong Kong government 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 increases.

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."

1. **REVIEW AND DATA ANALYSIS 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:

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

5.2 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.

5.3 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.

5.4 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.

5.5 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.

5.6 Descriptive Data: Some columns contained detailed descriptions of the properties in object data form, which might not be directly useful for the analysis.

5.7 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).

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

5.9 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.

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

**6. IDENTIFICATION OF THE TARGETS AND THE VARIABLES**

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

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

* 1. For each property transaction there was a unit rate of the property (in HK$ per square foot) (the unit\_rate); and
  2. For each property transaction there was the year-on-year price increase (called “PI”) based on the quarter the property was sold.

The research suggested that the 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 increase of the property could be a result of other macro economic factors.

For this reason, it was decided that in analyzing residential property prices and price increases. This analysis would be done on the following three targets:

* 1. The property price at a specific time (this was called the “new\_price” for reason set out below);
  2. The unit rate of the property in HK$ per square foot (which was called unit\_rate); and
  3. The price increase (called “PI”) of the property.

These factors were designated targets used in the analysis, with each target assessed by the machine learning algorithms separately.

1. **DATA CLEANING AND MANIPULATION**

7.1 Data Cleaning

As set ou above, several issues were identified with the merged\_dataset which was planned to be used for the machine learning analysis. This resulted in a number of sets to clean the date before analysis.

The property transactions in the dataset appeared to be comprehensive and record all property sales over a brief period of three years in reality it was incomplete and inconsistent for all districts.

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 the descriptive data was often left blank. Given this inconsistency, these fields needed to be removed from the dataset.

Other information such as sea view, distance to schools, distance to public transport are was missing from the property transactions dataset, which, considering the research identified in the literature review, could have also influenced property prices.

As discussed above, the information within the property transactions dataset varied significantly across different districts, with some districts being well-represented while others were not. Also, in the initial years (2020 and 2021), not all property transactions seemed to have been inputted or recorded into the property transactions dataset.

Additionally, to facilitate analysis using machine learning methodologies, all 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. 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 dataset with each sales transaction.

* 1. 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 folloowing techniques:

* + 1. 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 transactions dataset property are aligned with the relevant quarters of the macroeconomic data.
    2. 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.
    3. 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.
    4. Data Type Conversion: 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 data sets.
    5. 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.” Considering these descriptions, they appeared to be inconsistent and irrelevant to all of the properties, and these columns were deleted, eradicating the problems of null values.
    6. The Problem with the Prices. The price column in the property transactions dataset needed to be corrected. This was because it was the price values were rounded as it did not reflect the area multiplied by the unit rate per square foot. For this reason, the actual prices were determined using area multiplied by the unit rate to ensure accuracy. In the merged dataset, this new column with a corrected price called “new\_price” was created, and the old column called “price” was deleted or dropped.
    7. 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.
    8. 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 was used to scale the data between 0 and 1.
    9. One Hot Encoding for the Districts. One of the most important predictors of property prices is the area/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.

1. **FEATURE ENGINEERING**

Based on the dated in the merged\_database it was also possible to create new features and the following new features were created which could then be used in the analysis. These were:

* + 1. One Hot Encoding for the Districts. The research suggested that of the most important predictors of property prices is the area/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.
    2. Creating regions from districts. The merged\_dataset contained 18 districts and it was felt that that it might be useful to groups these districts into 3 three regions such as Island District, New Territories East and New Territories West.
    3. Classifying flat size. The merged\_dataset contained all different sizes of property from large to small. Property in Hong Kong is small by international standards and it was felt to would be useful group properties by size such as “small” (less that 500 square foot), “medium” (500 to 1000 square foot) and “large” (greater than 1000 square foot).

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

1. **CORRELATION ANALYSIS**

Before carrying out any machine learning analysis, it was necessary to discover useful information about the merged\_dataset in order to understand the data.

As the objective was to consider the factors which influenced the property price or price increase correlation with other variables would be important.

A correlation analysis using a heat map was performed which visualizes correlations between the three specific targets[[10]](#footnote-10) 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..
  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 Increase (“PI”) 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 price increases. 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 increases.

The weak correlations of unit rate and new price with macroeconomic indices are contrasted with the strong correlations of the price increase 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).

**10. 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:

* 1. 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.
  2. 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.
  3. Price Increase: 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 perform well for the `PI` (Price Increase), 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 increases 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.

1. **ANALYSIS OF PROPERTY PRICES BY DISTRICTS AND REGIONS**

The objective of this project was to analyze the factors affecting property prices and their changes. Therefore, it was crucial to examine property prices over the years and identify which areas (or districts) of Hong Kong experienced price changes. Additionally, it was important to assess whether the data for each district was complete for the entire period.

To achieve this, a data analysis was conducted to review the average prices for each district from 2020 to 2023. However, this analysis faced significant challenges due to the incomplete data for all 18 districts, especially for the early period of 2020 to 2022.

The incompleteness of the data presented several problems. There was limited data for some districts and certain yearly quarters, and scant information regarding property sales in the early period for many districts. This highlighted the data's lack of completeness.

There was no clear explanation for the incomplete data. It might be due to the sales offices in each district not updating their databases/forms until later in the recording period, or it could be attributed to the impact of the Covid-19 pandemic, during which fewer property sales might have been recorded, or the documentation process might have been neglected.

It is difficult to reconcile the notion of zero property sales 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 forms/database in the early and or Civid period.

Due to these issues, the in missing data in the database for all districts it would likely skew the analysis of the merged dataset. Despite this machine learning was applied the data.

1. **MACHINE LEARNING APPROACHES AND 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.

* 1. Regression Analysis;[[11]](#footnote-11)

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:

* + 1. 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.
    2. 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.
    3. Price increases (PI). The regression analysis explains that 99.8% of the variance in price Increase is 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 increase variance, reinforcing their importance in price increases.

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 increases.

12.2 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:

12.2.1. Unit Rate. The `unit\_rate` model shows moderate performance with an MSE[[12]](#footnote-12) and a R2 score. [[13]](#footnote-13)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.

12.2.2. New Price. For `new\_price`, the model's performance is also shows a moderate MSE and R2 score[[14]](#footnote-14) 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.

12.2.3 Price Increases. The model demonstrates excellent performance with an MSE and a R2 score of 1.0,[[15]](#footnote-15) with the graph showing a near-perfect alignment. This suggests that the features accurately capture the relationship between factors and price increases, 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.

12.3 XGBoost; [[16]](#footnote-16)

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:

* + 1. 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.
    2. 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:

* 1. For all three targets of price, unit rate and price increase all showed the best results using 3-fold cross-validation with a total of 729 fits.
  2. For price the best tuning was: `max\_depth=9`, `eta=0.1`, and `subsample=1.0`, yielding an R-squared of 0.984;
  3. 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.
  4. For Price Increase 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:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Metric | Unit Rate | | Price | | Price Increase | |
|  | 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.

* 1. MLP;[[17]](#footnote-17)

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

* + 1. 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`.
    2. Pice. 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.
    3. Price increase. 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.

1. **COMPARISON OF RESULT BETWEEN EACH MACHINE LEARNING MODEL**

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

|  |  |  |  |
| --- | --- | --- | --- |
|  | Regression Analysis | | |
| Target | Unit Rate | Price | Price Increase |
| 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 Increase |
| 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 increases |
| 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 Increase |
| 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 Increase |
| 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 Increases, suggesting feature engineering. XGBoost demonstrates effective learning and outstanding performance for Price Increase, 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 Increase predictions, while Decision Trees effectively capture this relationship when features are well-aligned.

1. **CHALLENGES AND PROBLEMS**

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[[18]](#footnote-18) also had their own limitations, further complicating the analysis.

1. **CONCLUSIONS AND 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.

1. **FUTURE RESREACH AND FINAL 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.

**REFERENCES**

1. The student is a retired director of the law firm Pinsent Masons and is a litigation solicitor, arbitrator, mediator, and adjudicator in the HKSAR, and an MSc Student at the University of Hull; Tim invests in property in Hong Kong, Thailand, and the UK. https://www.linkedin.com/in/timhallworth?utm\_source=share&utm\_campaign=share\_via&utm\_content=profile&utm\_medium=ios\_app. [↑](#footnote-ref-1)
2. The following were used: Decision Trees, Regression, XGBoost (Extreme Gradient Boosting), and MLP (Multi-Layer Perception) [↑](#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. The heads of the two dataset are set out in Appendix A [↑](#footnote-ref-6)
7. See Appendix 1 [↑](#footnote-ref-7)
8. https://www.kaggle.com/datasets/cyrusttf/hong-kong-housing-price-2020-2023 [↑](#footnote-ref-8)
9. See Appendix 2 [↑](#footnote-ref-9)
10. The target columns were `unit\_rate`, `PI`, `new\_price` [↑](#footnote-ref-10)
11. Regression analysis in machine learning models the relationship between a dependent variable and one or more independent variables to predict outcomes. The results provide coefficients that show how much the dependent variable changes with a one-unit change in each independent variable, the intercept indicating the expected value when all predictors are zero, and the R-squared value that measures how well the model explains the variance in the dependent variable. Additionally, p-values assess the significance of each coefficient, while residuals (differences between observed and predicted values) help evaluate model accuracy and identify potential patterns or biases. [↑](#footnote-ref-11)
12. The MSE score measures the average squared differences between actual and predicted values. MSE for the `unit\_rate` model is 41,282,685.68, indicating substantial prediction errors, as the average squared difference between actual and predicted values is high. [↑](#footnote-ref-12)
13. The R-squared (R²) score of 0.4827 for the `unit\_rate` model indicates that 48.27% of the variance in `unit\_rate` is explained by the model. This suggests the model captures some factors but leaves 51.73% unexplained, highlighting the potential need for additional features or more complex modelling. [↑](#footnote-ref-13)
14. The `new\_price` model has an MSE of 108,670,131,892.70, indicating large prediction errors. The R-squared (R²) of 0.43 means 43.08% of the variance in `new\_price` is explained by the model, suggesting that the model and features are insufficient for accurate predictions, leaving 56.92% unexplained. [↑](#footnote-ref-14)
15. An R-squared (R²) of 1.0 means the model explains 100% of the variance in `PI`, resulting in perfectly accurate predictions. [↑](#footnote-ref-15)
16. XGBoost (Extreme Gradient Boosting) is a machine learning algorithm based on gradient boosting that builds an ensemble of decision trees sequentially, where each tree corrects errors from the previous ones. The aim is to include regularization to prevent overfitting, handling missing values, and parallel processing for faster computation. [↑](#footnote-ref-16)
17. MLP (Multi-Layer Perceptron) models should be effective for forecasting new\_price, PI, and unit\_rate due to their ability to capture non-linear relationships and handle complex datasets. It offers flexibility and scalability, making them suitable for diverse forecasting tasks. Overfitting is a risk, necessitating regularization and cross-validation. Despite being less interpretable, MLPs provide robust, accurate predictions when properly managed, making them valuable for time series forecasting and other predictive tasks. [↑](#footnote-ref-17)
18. https://github.com/just4jc/HKProp/blob/master/HKProp\_Dataset.csv [↑](#footnote-ref-18)