DATA MINING PROJECT

Mining in the Property Market

OBJECTIVES

To develop an innovative data mining tool on the property market

TAM KAI TIK

22239065

Contents

1. Introduction

1.1 Background and Objectives

The task of this project is to develop an innovative data mining tool on which to base your startup business as an entrepreneur. Given that I am a business student, I decided to choose education, finance, and housing in an open-source centre in government.

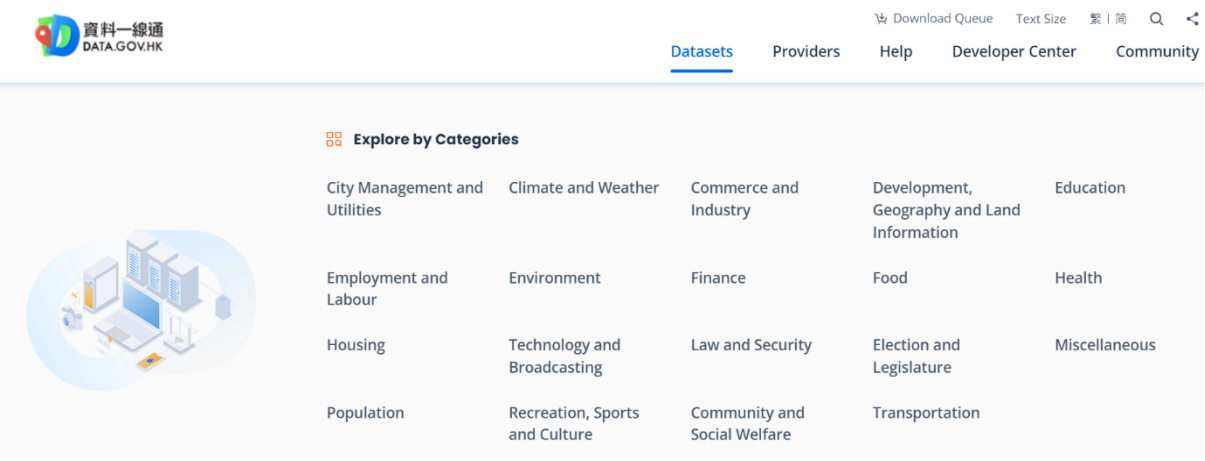

Figure 1 Categories of open source centre

After I read through the dataset in each category, the majority of the dataset only contains a few data or could not link to other datasets, at least I am seeking the dataset I could not image the usage of that dataset. Eventually, I chose the dataset inside the housing categories.

Housing is a fundamental necessity for human lives. As we all know, housing price is constantly increasing. It is believed that if you are trying to buy your own house further, its price would increase. Hence, as it is necessary and the price is high, it is needed to analyze the market and timing to buy a house. Given that the price is high, most of us might regret this once change for buying a house if we bought it at the wrong time such as buying at the peak price, it is hard to resell the apartment. In addition, conflicting news creates confusion. For example, it is easy to see the news says that the housing price is lower than the past quarter, which is challenging for Hong Kong economy. Hence, it is needed to uncover the pattern and do a data-driven decision-making for the general. And it is the objective or task of this project.

1.2 Target Audience

Given that housing is a necessity, the target audience is investors, general people, policymakers, or anyone who is interested in the Hong Kong property market.

In many housing markets, rents and home prices have been rising at a faster pace than incomes, hence it has prompted many to seek lower-priced options in market trends.

Saving a sizeable down payment is a major hurdle, especially for young buyers, hence it has prompted many to plan the mortgage for the compound interest calculator.

Beyond the purchase price, buyers face substantial additional expenses, including realtor fees, taxes, and mortgage interest. These hidden costs can add up quickly and pose a significant financial burden, especially for first-time homebuyers. It is crucial to uncover all the hidden expenses.

1.3 Methodology

First, it would develop an accurate and reliable forecasting model for property prices. By utilizing the data mining tool, it would uncover the pattern in the underlying dataset. Second, it would enable data-driven decision-making for investors, general people, and policymakers. Lastly, it would create a website that to visualize the model and make a business public.

2. Data Description

2.1 Data Sources and Data Description

The data sources are from the open-source centre in the government. The data providers are mainly in the Rating and Valuation Department. The collection of datasets is Average Rents by class monthly and yearly, GDP by year, Population by year, Class A (Vacancy) – Unit and Class A (Stock).

```
df.head()
```

	Date	Class A Hong Kong	Class A Kowloon	Class A New Territories	Class B Hong Kong	Class B Kowloon	Class B New Territories	Class C Hong Kong
0	1999-01-01	190	171	133	199	165	118	249
1	1999-01-02	196	173	133	204	165	114	239
2	1999-01-03	199	170	133	197	160	117	247
3	1999-01-04	191	171	135	200	156	116	256
4	1999-01-05	191	175	127	188	155	113	233

Figure 2 First 5 rows of the dataset

In this report, it is mainly focused on Class A, in which the saleable are less than 40m2. Additionally, the rent indicator by class is from 1999 to 2023 which resizes the time frame of the dataset to 1999 to 2023.

3. Data Preprocessing

3.1 Data Cleaning and Preparation

Given that the majority of the dataset is from the Hong Kong government, it is not user-friendly for data miners. It is used for report purposes. Hence, it is needed to modify the dataset so that we could use the dataset in Python. In addition, it is also needed to change the datatype and remove all duplicates or Null values.

```
[1]: import pandas as pd
     import matplotlib.pyplot as plt
     from statsmodels.tsa.seasonal import seasonal_decompose
     from statsmodels.tsa.arima.model import ARIMA

     # Convert data to Pandas DataFrame
     col_names = ['Date', 'Class A Hong Kong', 'Class A Kowloon', 'Class A New Territories',
                  'Class B Hong Kong', 'Class B Kowloon', 'Class B New Territories',
                  'Class C Hong Kong', 'Class C Kowloon', 'Class C New Territories',
                  'Class D Hong Kong', 'Class D Kowloon', 'Class D New Territories',
                  'Class E Hong Kong', 'Class E Kowloon', 'Class E New Territories']

     df = pd.read_csv("1.1M.csv", names=col_names, header=0)

     # Convert the first column (Date) to datetime format
     df[df.columns[0]] = pd.to_datetime(df[df.columns[0]])

     # Check the data type of the first column
     df.iloc[:, 0].dtype

[1]: dtype('<M8[ns]')
```

Figure 3 Examples of data cleaning in Python

4. Data Mining

4.1 Decision Tree Model

A decision tree model was built to predict the "Class A Hong Kong" property rental prices by using GDP, population, stock, and vacancy. Each internal node in the tree represents a decision based on an attribute, while each leaf node represents the outcome or class label.

```
# Select the target variable and features
X = merged_df[['GDP', 'All age group population', 'Class A Hong Kong difference', 'Class A (Vacancy) - Unit', 'Class A (Stock)']]
y = merged_df['Class A Hong Kong']
```

Figure 4 Target variable of the decision tree model

The model can explain approximately 71.96% of the variance in the target variable ("Class A Hong Kong"). However, it still has 28% of the variance in the target variable unexplained by the model.

```
Mean Squared Error: 2897.8
R-squared: 0.7196227320577125
```

Figure 5 Result of the decision tree model

The decision tree model can be visualized as a tree-like structure, where each internal node represents a feature, each branch represents a decision rule, and each leaf node represents a prediction or outcome.

The visualization shows that the "population" feature appears to be the most important predictor. It is because it is used as the first decision node that means the "population" plays a crucial role in predicting property rental prices. Secondly,

This suggests that population plays a crucial role in determining property rental prices. Other important features used in the subsequent decision nodes include "Class A (Stock)," indicating the potential influence of housing supply on rental prices.

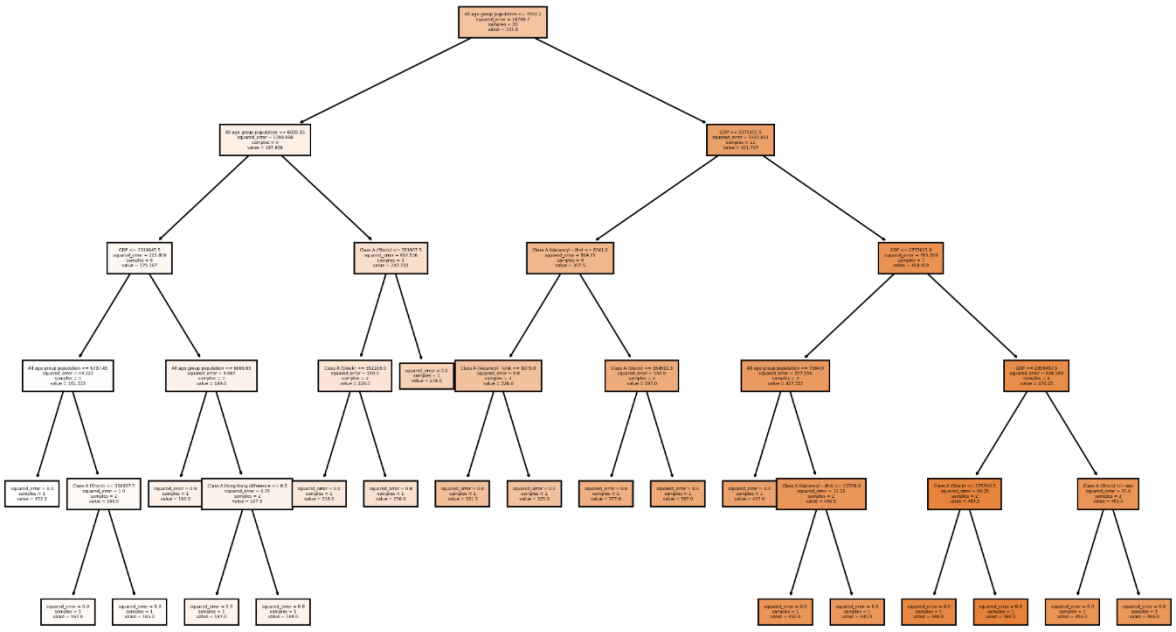

Figure 6 decision tree model visualization

4.2 LSTM (Long Short-Term Memory) Model

LSTM model was employed for forecasting property rental prices. Compared to decision tree, LSTM model offers several advantages to forecast. First, it is the ability to model time series data. In data mining course, neural networks with logistic regression and sigmoid function are to classify binary data. As LSTM allows to process time series data, it is good for us to forecast the rental prices. Second, LSTM allows new incoming data which able to adjust to changing trends and patterns. In full linked neural networks, it uses backpropagation to adjust parameters in the training data. It is not effective for incoming data. Hence, LSTM is decided to do the forecasting tasks. In addition, LSTM could capture long-term dependencies and patterns (e.g. seasonal patterns). In forecasting task, it also might uncover the pattern instead of using a linear regression model to predict the next month's prices. In the linear regression model, it was tried but the result was bad hence the model changed to LSTM.

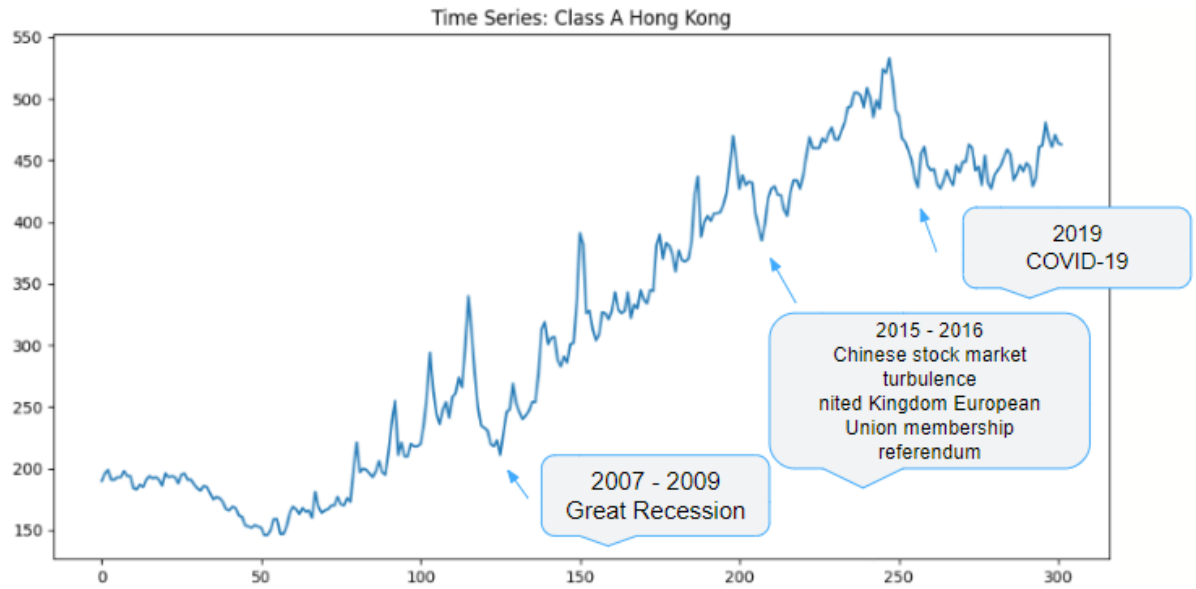

Figure 7 Time Series Plot of Class A Hong Kong

It is the time series plot of class A Hong Kong. As the biggest decrease is related to Stock market crashes in Hong Kong, it might add it into the model to train. However, as the task of the model is to forecast the rental price, when the stock market crash appears, it is too late to add it into the model. Also, the diagram trend is remains increasing. hence, it is proved that the news or general perspectives of the property market are true.

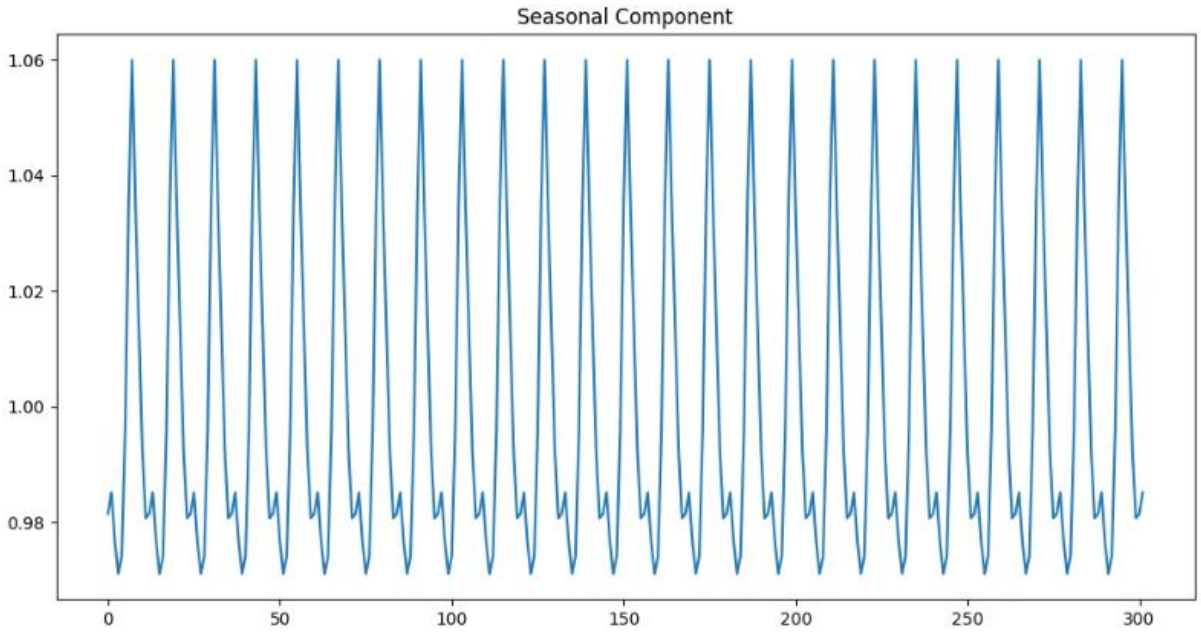

Figure 8 Seasonal component in LSTM model

This plot represents the seasonal component in LSTM model. The y-axis of the graph shows the seasonal component values. The values greater than 1.0 which is a positive seasonal effect otherwise lower than 1.0 which is negative. If it is a negative seasonal effect that means the observed values tend to be lower than the overall trend. In this graph, we could observe a distinct seasonal pattern that repeats every 12 time periods.

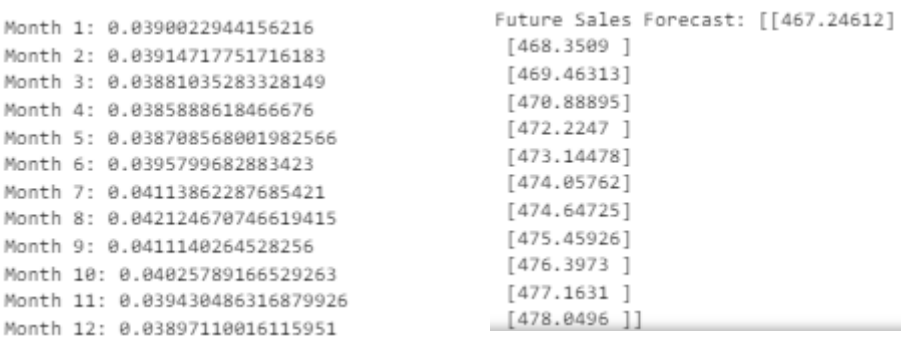

Figure 9 Result in LSTM model

Those values correspond to seasonal coefficients for months. The highest value is 0.04212467 in August. The lowest value is 0.03858886 in April. Hence, for people who want to buy a house, it is good to make a decision in April.

Product Selection

5.1 Website for Property Problem

Given that LSTM and Decision model are developed by historical data in the property market and other general economic data namely GDP. By applying data mining techniques, decision tree could explain 72% of the variance in Class A rents and LSTM could capture the series pattern in Class A. To deliver these insights effectively, this project is going to deploy a website integrated with various features on it.

5.2 Features

As the website is not deployed by WordPress, the initial cost would be the domain fee which is USD10 per year. Assumed as an entrepreneur, a lower cost would be an advantage to survive in the market. Also, it integrates different data sources such as government, Bank. By using amChart on the website, it improves data visualization and user experiences. Compared to the popular property website, it is user-friendly and it is an All-in-One product. Although the real estate valuation is still done by the Bank, it is good to put all information in a single website. For the return or business model in this project, it would integrate agent collaboration, value-added services and influencer marketing services. After using above service, the service fee would be charged. Also, it would build up a community feature namely forums and discussion. By increasing the reach of the website, it would earn a little money from Google AdSense. It could handle the domain fee. Hence, it is a lost-cost project but has a high return. In addition, if the result of this project is good, it could minus the project to all countries. Hence, it has a great potential market value especially since population is increasing, and housing is a necessity.

Conclusion

In conclusion, the data mining project aimed to develop an innovative tool for analyzing the Hong Kong property market by using various data sources. By applying data mining techniques, decision tree could explain 72% of the variance in Class A rents and LSTM could capture the series pattern in Class A.

To effectively spread these findings and enable data-driven decision-making for investors, the general public, and policymakers, it is integrated by all data sources and improved the past way to buy a house. And the website would be the way to represent the findings in this project.

References

1. *Home | DATA.GOV.HK*. (n.d.). https://data.gov.hk/en/