MATH3836 Data Mining

Project in Private Domestic Statics

22239065 TAM Kai Tik

Agenda

1. Introduction and Background
2. Data Loading and Preprocessing
3. Model Selection
4. Model Visualization
5.

Introduction and Background

Topic Selection

資料一線通
DATA.GOV.HK

Download Queue | Text Size | 繁 | 简

Datasets | Providers | Help | Developer Center | Community

Explore by Categories

City Management and Utilities	Climate and Weather	Commerce and Industry	Development, Geography and Land Information	Education
Employment and Labour	Environment	Finance	Food	Health
Housing	Technology and Broadcasting	Law and Security	Election and Legislature	Miscellaneous
Population	Recreation, Sports and Culture	Community and Social Welfare	Transportation	

Introduction and Background

Problem Identification and Objectives

- Housing is a **fundamental necessity**
- Housing prices constantly **increasing**
- Conflicting news create **confusion**

Objectives

- Develop an accurate and reliable forecasting model for property prices

- Enable data-driven decision-making for investors, general people, and policymakers

Introduction and Background

Target Audience

High Housing Costs

- Rents and home prices rising faster than incomes in many markets -> seeking for a lower price in market trend

Down Payment Obstacles

- Saving sizeable down payment a major hurdle, especially for young buyers -> planning for the compound interest

Significant Transaction Costs

- Realtor fees, taxes, mortgage interest add major expenses -> uncover all the hidden expenses

Introduction and Background

Data Description

Data Providers:

Rating and Valuation Department

Dataset:

Property Market Statistics |
Private Domestic - Average Rents by Class - Monthly (from 1999)

Remarks:

Price and Rental Indices

*Annual Rent / Rateable Value
(Annual Rental Value assessed by the government)

```
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

Introduction and Background

Data Description

Further Dataset:

1. GDP by year
2. Population by year
3. Class A (Vacancy) - Unit
4. Class A (Stock)

Private Domestic units are defined as independent dwellings with exclusive cooking facilities, bathroom and toilet. They are classified by reference to floor area as follows:

Class A - saleable area less than 40 m^2
Class B - saleable area of 40 m^2 to 69.9 m^2
Class C - saleable area of 70 m^2 to 99.9 m^2
Class D - saleable area of 100 m^2 to 159.9 m^2
Class E - saleable area of 160 m^2 or above

Remarks:

Given that Rents indicator is from 1999 to 2023, it would use this time frame to forecast.

Data Preprocessing

```
[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]')
```

Model Selection - Decision Tree

```
# 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']
```

R-squared (R^2):
The model can explain approximately 71.96% of the variance in the target variable ("Class A Hong Kong")

It still have 28% of the variance in the target variable unexplained by the model

Model Result

```
R-squared: 0.7196227320577125
```

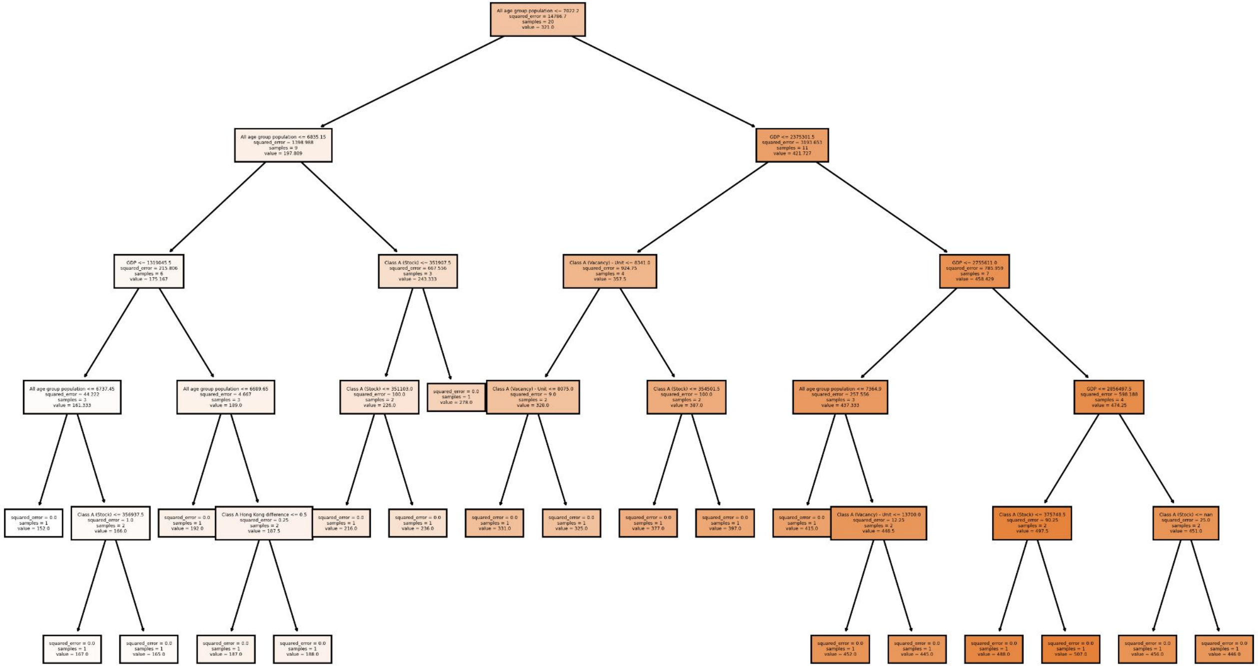
Model Visualization

Model Visualization – Decision Tree

Example:

1. population <= 7022.2
2. population <= 6835.15
3. Class A (Stock) <= 351907.5
4. if Class A (Stock) > 351103.0

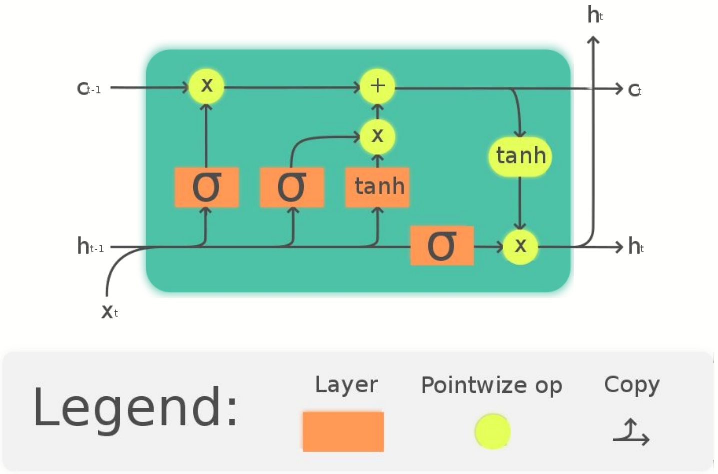
-> 278 (Forecast Rent this year)



Model Selection - LSTM (Long Short-Term Memory)

Benefits:

- Ability to model in time series data
- Ability to adapt to new data
- Capability to capture long-term dependencies and patterns (e.g. seasonal pattern)
- Effective in handling sequential data with trends and seasonality
- Proven success in various time series forecasting applications (e.g. stock market prediction, sales forecasting)



Model Visualization

Building the LSTM Model

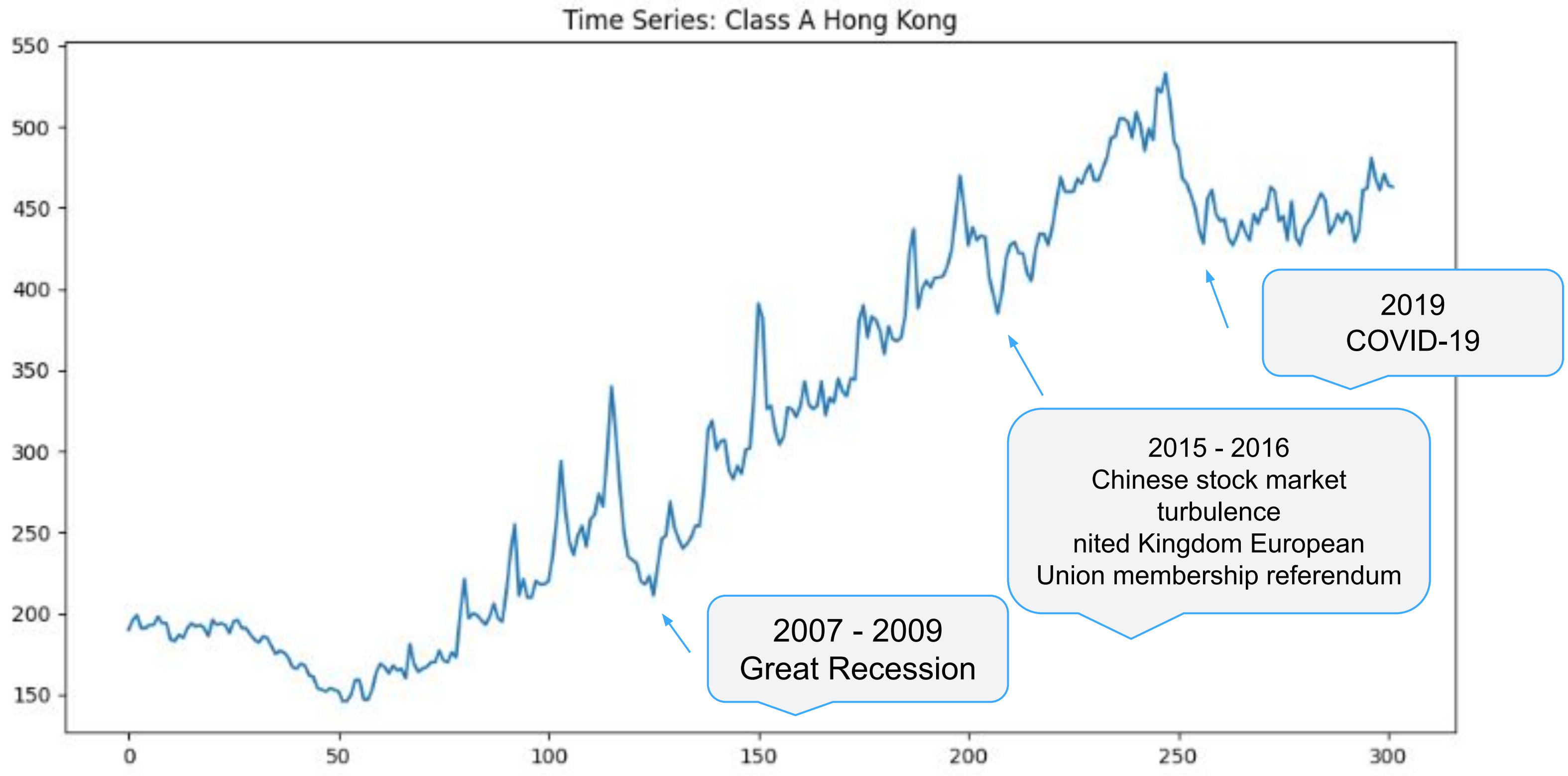
```
# Select the category for analysis
category = 'Class A Hong Kong'
series = df[category].values
# Plot the time series
plt.figure(figsize=(12, 6))
plt.plot(series)
plt.title(f'Time Series: {category}')
plt.show()

# Perform seasonal decomposition
result = seasonal_decompose(series, model='multiplicative', period=12)
result.plot()
plt.show()

# Train ARIMA model and forecast for the next 12 months
model = ARIMA(series, order=(2, 1, 2))
model_fit = model.fit()
forecast = model_fit.forecast(steps=12)
print(forecast)
```

Model Visualization

Data Visualization



Model Visualization

Data Visualization

1. Time Series Plot:

- The overall trend in property prices over time

2. Seasonal Decomposition Plot:

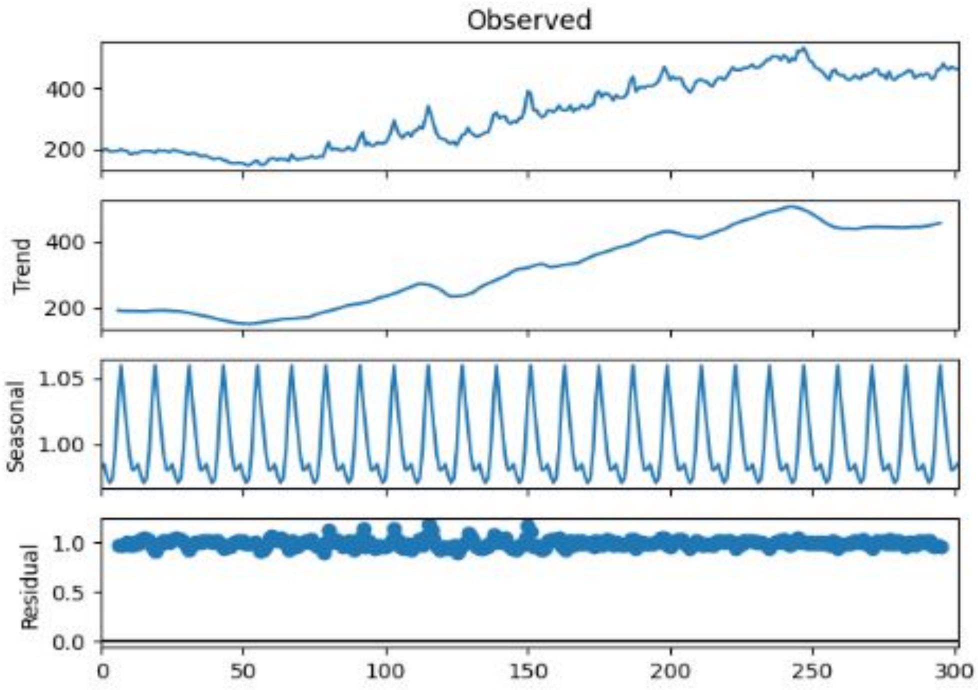
- Trend: The underlying long-term trend in the data

- Seasonal: The recurring seasonal patterns

- Residual: The remaining variations not captured by the trend or seasonal components

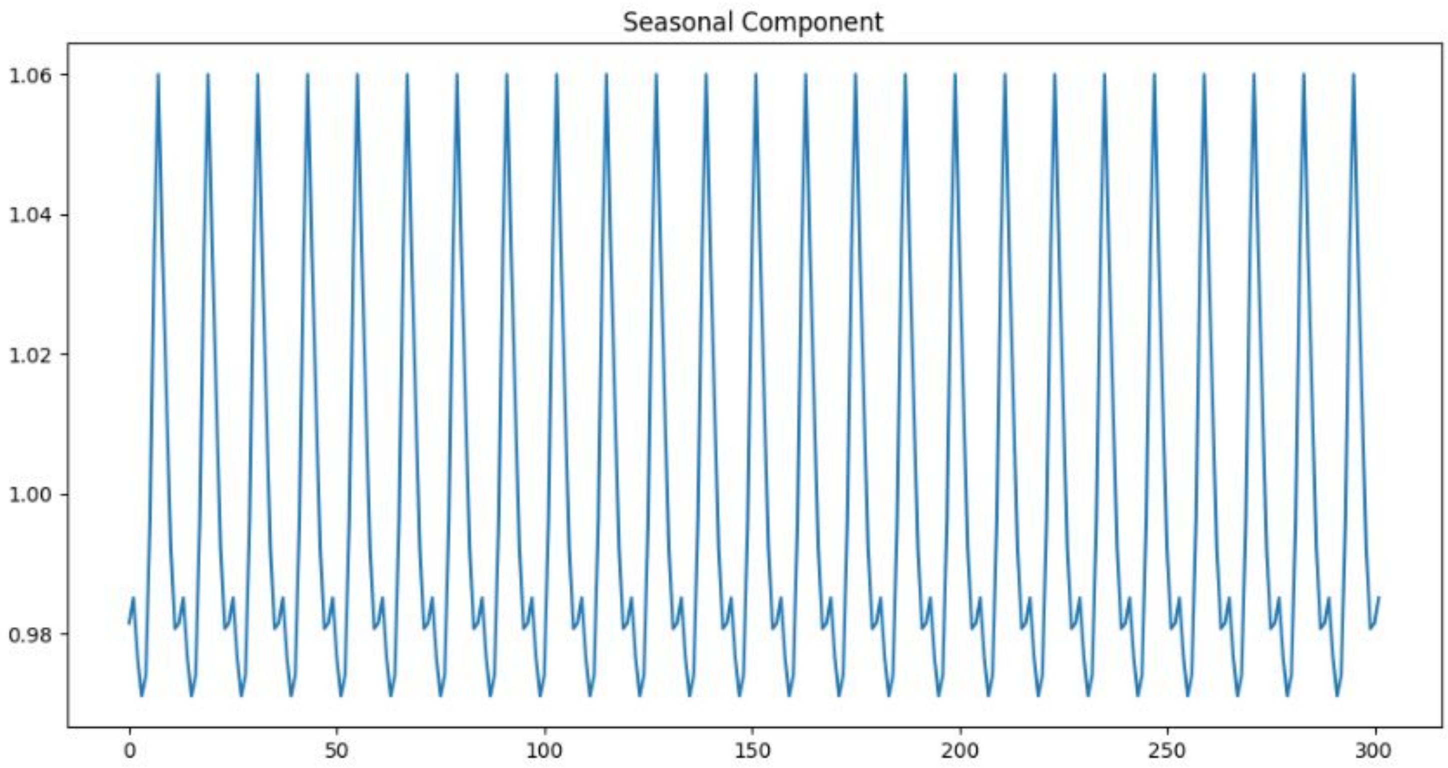
3. ARIMA Model Forecast:

The code trains an ARIMA (Autoregressive Integrated Moving Average) model on the time series data and generates a forecast for the next 12 months




```
[463.28729174 463.20178565 463.14743976 463.1192673  463.10490096
 463.09758586 463.09386164 463.09196561 463.09100033 463.0905089
 463.0902587  463.09013133]
```

Model Visualization
Data Visualization



Model Visualization

Data Visualization

Values correspond to seasonal coefficients for Jan, Feb, Mar...Dec

- Greater than 1 indicates months tend higher than overall trend, otherwise tend lower than overall trend

Highest value 0.04212467 for August

- Observations tend higher than trend in August

Lowest value 0.03858886 for April

- Observations tend lower than trend in April

```
Month 1: 0.039002294156216
Month 2: 0.03914717751716183
Month 3: 0.03881035283328149
Month 4: 0.0385888618466676
Month 5: 0.038708568001982566
Month 6: 0.0395799682883423
Month 7: 0.04113862287685421
Month 8: 0.042124670746619415
Month 9: 0.04111402645282 56
Month 10: 0.0402578916652926 3
Month 11: 0.039430486316879926
Month 12: 0.03897110016115951
```

Product

Product Selection

A website for Property Problem (https://math-3836-data-mining-project.vercel.app/)

Reason:

1. Cost-effectives (one domain with in USD 10 per year)
2. Integrated Data Sources
3. Improved Data Visualization and user experiences
4. Integrate Agent Collaboration (Service fee after render a buyer)
5. Offer Value-Added Services (mortgage calculations, agents help)
6. Leverage Influencer Marketing (Service fee after render a buyer)
7. Implement Community Features (forums, discussion)