



**City University of Hong Kong
Department of Computer Science**

BSCCS Final Year Project Report 2015-2016

(15CS103)

**Big Data Analysis on Housing Price Index from
mass media, government data and the real estate data**

(Volume 1 of 1)

Student Name : LAM Chi Ho

Student No. :

Programme : CS4514
Code

Supervisor : Dr. WONG, Ka Chun

1st Reader : NUTANONG, Sarana

2nd Reader : ZHANG, Qingfu

For Official Use Only

Student Final Year Project Declaration

I have read the project guidelines and I understand the meaning of academic dishonesty, in particular plagiarism and collusion. I hereby declare that the work I submitted for my final year project, entitled:

Big Data Analysis on Housing Price Index
from mass media, government data and the real estate data

does not involve academic dishonesty. I give permission for my final year project work to be electronically scanned and if found to involve academic dishonesty, I am aware of the consequences as stated in the Project Guidelines.

Student
Name: LAM Chi Ho Signature: _____

Student ID: _____ Date: _____

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Dr. WONG, Ka Chun. He has provided advice and suggestion for me in the final year project.

Also, I would like to thank Jamie Wan, who gave so much useful advice and mental support to me.

Last but not least, I would like to thank my parents for their kindly encouragement and show my appreciation to all of them. Without their support, it will be difficult for me to finish the report within limited time frame.

Abstract

Shelter is a basic human need. However, Hong Kong people are facing severe housing problem. Young generation could not afford to buy their own property by themselves owing to extremely high Hong Kong house price . Not only youngsters felt difficult to buy their house, even some people in Hong Kong was still living in such worse environment like “cage homes”, subdivided units and roof-top flat.

In recent years, Internet becomes more informative than that in the past since mass media and other companies are more likely to deliver their messages or information via website. In addition, Hong Kong government has established data.gov.hk to share open data to the public so as to utilize the usage of data and adopt the big data century, which allows the public to access and download their open resources.

This project aims to perform data analysis on Hong Kong housing problem making use of rich information from the Internet. Also, it provides an alternative way to investigate the housing issue. In this project, I applied data mining technologies to analyse data and perform data processing on data collected from the Internet.

Table of Content

Acknowledgement	3
Abstract	4
1. Introduction	8
1.1 Background Information	8
1.1.1 Unaffordable Housing prices	8
1.1.2 Unqualified Living Condition	9
1.1.3 Historical Housing Price Index	9
1.1.4 Data Anlysis	10
1.2 Project Objective	11
2. Literature Review	12
2.1 Data Mining Process	12
2.2 Fundamental Study on Housing in Hong Kong	13
2.2.1 Supply and Demand Chain	13
2.2.2 Intrinsic Value of Real Estate	13
2.2.3 Periodic Cycle of Real Estate Market	14
2.2.4 Relationship between Housing and Stock Market	14
2.3 Review on data mining techniques	15
2.3.1 Decision Tree with C4.5 Algorithm	15
2.3.2 Naïve Bayes Classification	17
2.3.3 k-Nearest Neighbors algorithm (kNN)	18
3. Data Collection and Preprocessing	19
3.1 Data Quality	19
3.2 Data Collection	19
3.2.1 Government Open Data	20
3.2.2 Web Scrapping	22
3.3 Data Preprocessing	23

3.3.1 Attribute Selection	23
3.3.2 Attribute Creation	23
3.3.3 Data Validation.....	23
4. Data Modeling	24
4.1 Transactional Data	24
4.2 Time Series Data.....	26
4.2.1 Dependent and Independent Variables.....	27
4.2.2 Multiple Linear Regression.....	27
4.2.3 Interpreting the correlation coefficient r	27
5. Model Evaluation.....	28
5.1 K-fold cross-validation	28
5.1.1 Reason of selecting 10-fold cross-validation.....	29
5.1.2 Error Rate	29
5.1.3 Confusion Matrix	30
5.1.4 Receiver Operating Characteristic (ROC) Curves.....	31
6. System Design	33
6.1 Use Case Diagram	33
6.2 Use Case Description.....	34
6.3 Sequence Diagram.....	35
6.4 Website Layout Prototype	37
7. System Implementation.....	38
7.1 Overview of System	38
7.2 Core Programming Language	38
7.3 Selected Platform and Tools	38
7.3.1 Waikato Environment for Knowledge Analysis (Weka)	38
7.3.2 Google App Engine (GAE)	39
7.3.3 Google Cloud Datastore.....	39

7.4 Web Application	40
7.4.1 Homepage	40
7.4.3 Data Access.....	41
7.4.4 Data Prediction	43
7.4.5 Data Visualization	44
8. Project Review and Conclusion.....	46
Appendix I – Monthly Progress Log	47
References.....	49

1. Introduction

1.1 Background Information

1.1.1 Unaffordable Housing prices

There is no doubt that home is a basic necessity of life. However, Hong Kong has severe housing problems for example unaffordable housing prices and rents, crowded and insecure living environment, which arouse the public attention and the request of addition of housing supply and construction. Hong Kong Chief Executive Officer Mr. Leung said that tackling housing problem should be the top priority of duty within his term of office. Despite the fact that Hong Kong government places much emphasis on handling housing issue, the housing policy launched could not have an instant and valid cool-down effect on hot housing market.

A survey of 360 cities by US-based Consultancy Demographic stated that Hong Kong ranks the first class for the most unaffordable housing for the fourth straight year among the world.[1] The survey also pointed out that the median cost of a flat in Hong Kong is almost 15 times the annual household income [1]. In reality, it is not feasible to save the total of 15 years of annual income without other expenditure in daily life, which means that the time needed to purchase a flat is even longer than 15 years in average. In the face of soaring property prices and rents, young generation in Hong Kong probably feel difficult and frustrated to purchase their own flat without their parents' financial support. Meanwhile, their salary income could not catch up the increase of property price in recent years, resulting in extreme financial pressure from purchasing housing, marriage and expenditures on bringing up children. Such pressure may also cause the social problem such as late marriage and not having children in family. The similar situation exists in both second-hand property market and rental market. Those two market is sensitive to the price change in first-hand property market. Therefore, it is a common phenomenon that youngsters live with their parents instead of purchasing or renting a flat in Hong Kong, even young couple after marriage live separately for a temporary solution of housing problem.

1.1.2 Unqualified Living Condition

Worse still, there are a small group of people living in so-called roof-top unit and subdivided unit. The subdivided unit only has a tightly space with poor air ventilation. With the removal of original structural partition walls and unauthorized construction of new toilets and kitchens, these unauthorized building works pose a potential danger of their safety. Yet they do not have much choices on their living environment. The government subsidized housing policy cannot instantly relieve their dilemma owing to the long waiting list for the scheme public rental housing (PRH). In accordance with the statistics of Hong Kong Housing Authority, the average waiting time for general PRH applicants was 3.7 years. It is important for the government to help all households in Hong Kong gain access to both affordable and appropriate housing service.

1.1.3 Historical Housing Price Index

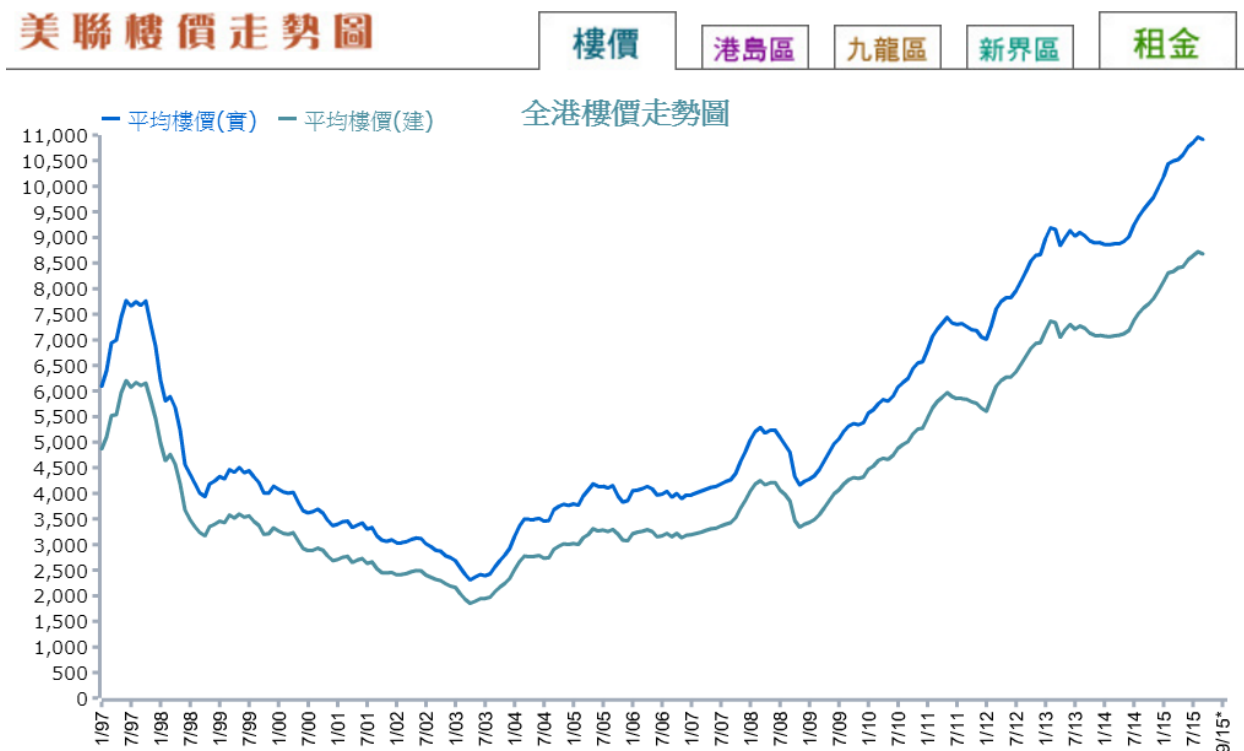


Figure 1 The Midland Market Property Price Chart from Jan, 1997 to Sept, 2015

In reference to the above line chart, the property price started dropping from the peak in 1997 because of 1997 Asian economic tsunami. An enormous financial crisis was impacting to those Asia countries and cities including Hong Kong. The 85,000 housing policy by former chief executives Mr. Tung. 85,000 public housing supplies each year amplified the damage on real estate market. The housing price was then dropping

rapidly from the peak at 1997. The decreasing trend continued and it reached its lowest point in 2003 owing to the break out of SARS. It then remained a gradual rise from 2004 to 2007. After the half quarter of 2009, there was an upward trend in property price. Even though the government strives to cool down the property market by supplying more public housing estates in recent years, it still reaches the highest point over the past 18 years. It reveals that Hong Kong government does not have a long term, well-planned and sustainable housing policy.

From the mentioned background information, a short summary to housing problem in Hong Kong is shown as follows.

- There are some Hong Kong people cannot acquire **affordable** and **qualified** living conditions
- Government plays an important role in stabilizing and addressing housing price in society when the property market goes extreme situation in neither high nor low. However, Hong Kong government probably lacks of a long-term and continuable planning on housing policy.

1.1.4 Data Analysis

The housing issue becomes more complicated than that in the past. It is the old saying that limited land resources and huge population led to the high real estate price. Nevertheless, it is not sufficient to explain the soaring increase in property market recently. In this project, it tries to figure out and estimate the factors affecting property market by taking account into the multiple factors like speculation in property market, government housing supply and stock market simulation. It is believed that the root of housing problem is not easy to identify due to many social and economic factors and hence a comprehensive analysis is required.

1.2 Project Objective

The objective of this project aims to perform data analysis on housing data for the sake of finding some useful knowledge and pattern.

It is divided into two major parts. The first part is to identify the major factors affecting residential property price in Hong Kong by examining stock stimulation, supply and demand in market, population and other relevant data. And the second part is related to the technical part of data analysis and it can be concluded in the following items:

1. To construct data analysis for processing housing related datasets from government, media and real estate company in order to
 - a. Discover hidden patterns or relationship
 - b. Perform prediction based on historical data
 - c. Present and organize the knowledge and statistics in a user-friendly way
2. To analyse the functionality of data analysis tools and make a comparison between them
3. To implement data cleaning for data collected from different sources
4. To study and select appropriate data mining algorithms
5. To construct a web application for users to access the service
6. To provide an alternative and attractive graphic interface to visualize the data instead of traditional tabular form of data

2. Literature Review

2.1 Data Mining Process

To perform data mining, it is necessary to review the process of data mining. Cross Industry Standard Process for Data Mining (CRISP-DM) is a widely used approach applied in the workflow of data analysis. It is decided to simplify the phases of CRISP-DM and extract its major characteristics into the process of data mining in this project. The simplified workflow is shown in Figure 2.

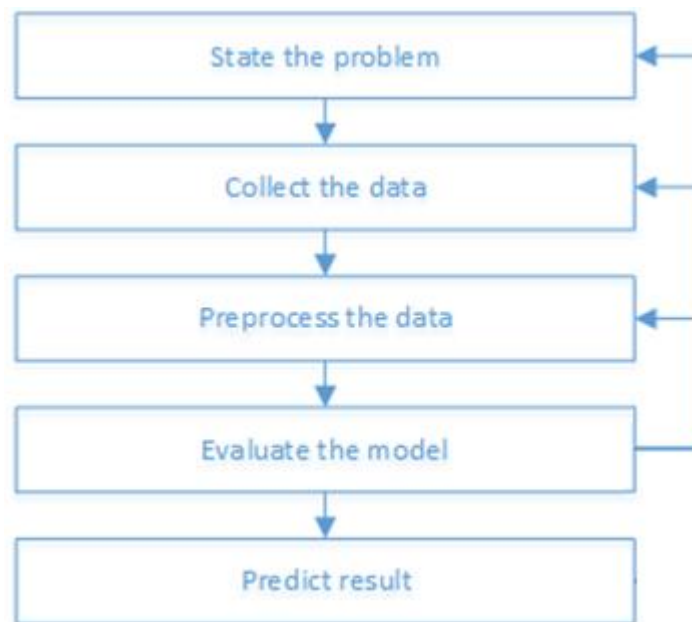


Figure 2 Data Mining Process

The first stage “State the problem” is completed in the previous section - background information and the project objective is pointed out in the section 1.2. The following part will explain further on the housing analysis in Hong Kong and figure out some clues on predicting housing price. After understanding the current situation, it is necessary to enter the stage of data collection. As there is no existing data warehouse or database, the data collection relies much on the information collected from government open data and the Internet.

2.2 Fundamental Study on Housing in Hong Kong

The first step of the process is to state the problem and understand the fundamental knowledge of Hong Kong housing. With the prior knowledge, it helps to construct data model and interpretation of result later.

2.2.1 Supply and Demand Chain

The supply and demand chain is of importance to determine the factors affecting housing price. When the demand is greater than the supply, the price goes up and vice versa. According to the report of “Housing in Figures 2015”, it shows that the type of housing can be divided into three major sectors, which are 30.4 percentages of public rental housing , 15.5 percentages of Home Ownership Scheme (HOS) under subsidized sale flats and 53.5 percentages of private permanent housing [2]. The owners of HOS and private permanent housing except PRH are eligible to sale their flat and those transaction data is recorded in government and real estate company websites. Apart from the available on-sale flats in the market, the government will supplement the residential land supply each year by launching the land scale to developers by auction and contruction of PRH and HOS. On the other hand, the demand is not only restricted to residential demand of own-occupiers, but also the investment need of potential purchasers. The residential demand depends on demographic factors like population growth, the formation of household and the number of marriage while the investment need depends on economic factors such as cosumer price index (CPI), gross domestic product (GDP) and stock market statistics.

2.2.2 Intrinsic Value of Real Estate

The intrinsic value of real estate refers to actual value regarding the property. For example, the intrinsic value includes age of building, location, flat size, floor etc. The intrinsic value of real estate is relatively stable. It should not be ignored when analyzing housing price. The intrinsic value can be well observed from the price difference between two flats at the same time. The flat with higher quality is more expensive than the other one.

2.2.3 Periodic Cycle of Real Estate Market

Like stock price, the housing price reflects the market behaviors so its moving trends in both increasing and decreasing is not a random movement. In other words, the housing price should be predictable. Henry George (1876) observed the curious cycle through which real estate market inexorably move: recovery, expansion, hypersupply and recession [3]. Such pattern of price index recurses over a certain period of time. In general, the periodic cycle of housing price is long, approximately more than five years for one complete cycle.

2.2.4 Relationship between Housing and Stock Market



Figure 3 Image from Property, stock indexes seen to hit new highs @ CHINADAILY ASIA

Many researches have been done to prove the relationship between housing and stock market. For instance, Woo and Chan concluded that real housing and stock prices have positive and two-way causal impacts on each other — evidence of wealth and credit-price effects.[4]. In the above figure, it shows the Hang Seng Index and Centa-City Index along the time axis. It is observable that the Hang Seng Index (HSI) and Centa-City Index (CCI) have almost the same bull market and bear market pattern with a little divergence.

2.3 Review on data mining techniques

Data mining is important in data analysis, which helps to discover association and hidden pattern in the data. It has a wide range of usages in computer science field including machine learning, optimization, data analysis etc. **Data mining techniques selected in the project** will be introduced from the aspect of underlying principle.

2.3.1 Decision Tree with C4.5 Algorithm

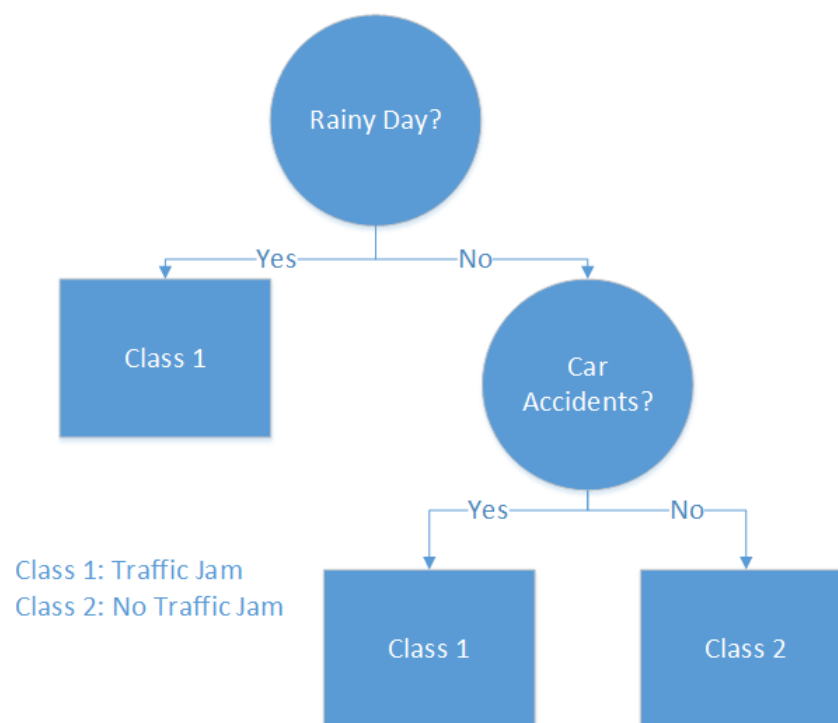


Figure 4 Simple Example of Decision Tree

Decision tree is a classic technique in classification algorithms for supervised learning. Similar to other classifiers, decision tree requires a set of input-output samples for constructing the model. It is a top-down recursive divide-and-conquer strategy searching the path from a starting root node down to one leaf node. Given in Figure 4, all samples with selected attribute value “Rainy Day: Yes” belong to class 1, the remaining samples will then enter the intermediate node in the next level. For example, a sample S with attribute (Rainy Day: No, Car Accidents: No) will be passed to the root node and check whether the attribute of S is rainy day or not. As the “Rainy Day” attribute value of S is “No”, then it will be passed to the next intermediate node for further classification. In that case, S will be classified as Class 2 because of its attribute

value of “Car Accidents: No”. Indeed, no further branch or intermediate decision node is required until the below conditions are achieved.

- The partitioned samples belong to the same class.
- There are no further attributes for classification.

The determination of attribute selection in C4.5 algorithm is the highest value of information gained by branching on attribute, which is calculated from the below formula:

Let A be the attribute and D be the set of all training samples. F(x) specifies the function of entropy.

$$Gain(A) = F(D) - \sum_{i=1}^v \frac{|D_i|}{|D|} \times F(D_i)$$

The decision tree is relatively simple and readable. However, it may be overfitting to the training data. Prunning is an important technique to discard one or more subtrees and replace them with leaves simplify a decision tree [5].

2.3.2 Naïve Bayes Classification

Naïve Bayes Classification is a conditional probabilistic model based on Bayes' Theorem.

Bayes' Theorem is stated as the following mathematical expression.

Let the training data be X and the posteriori probability of a hypothesis be H ,

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)}$$

Suppose there is a set of m samples $S = \{S_1, S_2, \dots, S_m\}$ where each sample is represented by n -dimensional attribute vector $X \{x_1, x_2, \dots, x_n\}$. Every sample belongs to one of k classes labeled as $\{C_1, C_2, \dots, C_k\}$. Using Bayes' theorem and substitute the variables, we get:

$$P(C_i|X) = \frac{P(X|C_i) \cdot P(C_i)}{P(X)}$$

$P(X)$ is constant for all classes, which implies the value is independent of C . Therefore, we only need to maximize

$$P(C_i|X) = P(X|C_i) \cdot P(C_i) \text{ and } P(C_i) = \frac{\text{number of samples in class } C_i}{m}$$

The assumption of naïve Bayes states that the features variables are conditionally independent of given class C . With the naïve assumption of conditional independence between attributes, the conditional probability can be decomposed as

$$P(X|C_i) = \prod_{k=1}^n P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \dots \times P(x_n|C_i)$$

To predict whether the sample belongs to which class, we only need to compare the probability of $P(X|C_i)$ among different classes. The sample with the highest value of $P(X|C_{\text{predicted}})$ will belongs to that class $C_{\text{predicted}}$ according to naïve Bayes classifier.

2.3.3 *k*-Nearest Neighbors algorithm (*k*NN)

*k*NN algorithm is a common data mining technique to classify the input. It determines the decision boundary based on the form of the distance function. *k*NN simply memorizes all samples in the training set and then compares the test sample with them [4]. It is assumed that if a test sample *X* is bounded in the local region *R*, *X* will be marked the same class label of *R*. The commonly used distance metric is Euclidean or Manhattan distance.

$$\text{Euclidean distance: } \sqrt{\sum_{i=1}^k (x_i - y_i)^2};$$

$$\text{Manhattan distance: } \sum_{i=1}^k |x_i - y_i|$$

To optimize *k*NN classification, it is vital to determine the value of *k*. With reference to Figure 5, the *k*NN classifier for *k* > 1 is more robust. Also, larger values of *k* reduce the effect of noise on the classification [6].

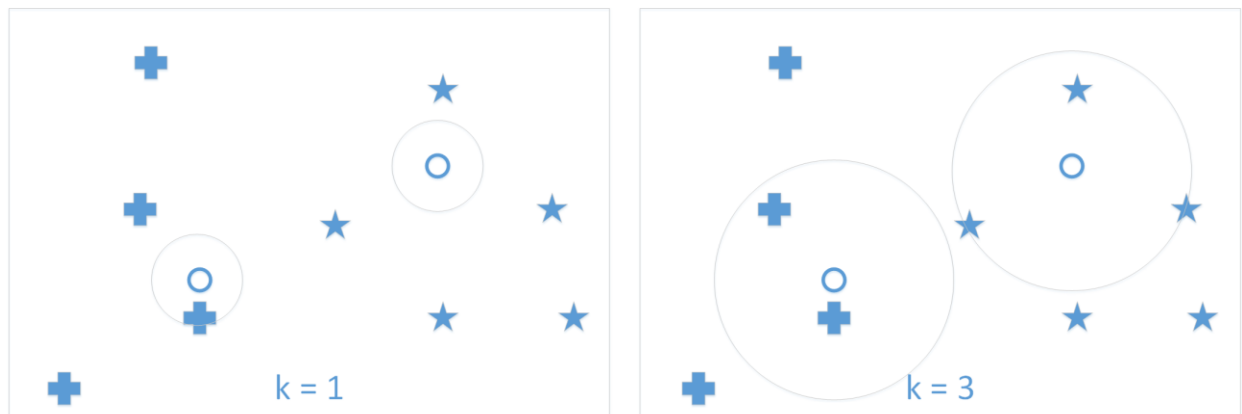


Figure 5 *k*-nearest neighbor classifier

3. Data Collection and Preprocessing

3.1 Data Quality

To acquire reliable data for making decision, data quality is an essential element when performing data collection and preprocessing. The quality of output depends on the quality of input as stated in the concept of “Garbage in, Garbage out”. Inaccurate information may lead to the failure of data interpretation and prediction. Therefore, the data should try to be clean, relevant and accurate.

3.2 Data Collection

Data collection is the process of gathering and generating information. To analyze housing problem in real life, it is more convincing to adopt the real data instead of the synthetic data for evaluation. Indeed, the web provides a large data repository in the world. The growth of Internet is incredibly fast. In 2013, Sintef found that a full 90 percent of all the data in the world has been generated over the last two years [7]. The trend of mobile technology contributes to the data volume in the Internet. Moreover, Internet provides a convenient platform for information sharing; thus, more companies and mass media are likely to deliver their messages and content through the web. In spite of the fact that much knowledge can be extracted from the Internet, it is challenging to filter out useful data from the noise and other irrelevant information.

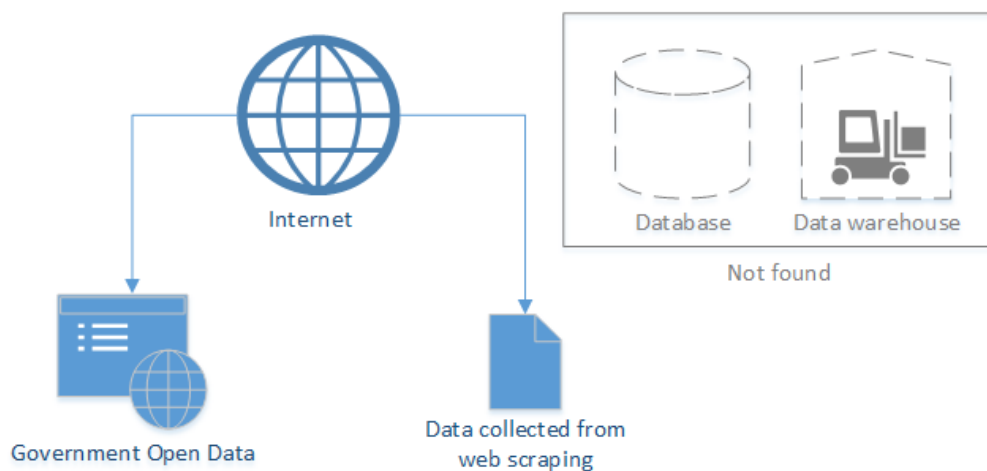


Figure 6 Illustration of Data Collection

3.2.1 Government Open Data



Figure 7 Screen Capture from data.gov.hk

On 18th March 2015, the Office of the Government Chief Information Officer (OGCIO) launched the public section information (PSI) portal data.gov.hk in support of the digital era. It is believed that providing PSI in a form or on terms that facilitates its wider dissemination and re-use will increase the value that the community realizes from the use of such information [8]. A well-designed open data webpage can utilize the usage of government in data analysis. However, the truth is that the data provided in data.gov.hk is not as user-friendly as expected. From table 1, it is observable that the volume and the availability of different data formats of data.gov.hk is less outstanding than the open data platform in other countries.

Table 1 Comparison of government open data platform

	Data.gov.hk (Hong Kong)	Data.gov (United States)	Data.gov.tw (Taiwan)
Number of datasets	Over 5,000	Over 1,784,864	Over 14,000
Search function	Yes	Yes	Yes
Support of multiple data format	Only support one specific data format	More than one option for data format sometime	Always more than one option for data format

During studying data collection from data.gov.hk, it is found that numerous data is stored in Excel format or portable document format (PDF), which is difficult to automate the data extraction owing to its irregular data scheme. Mr. Wong pointed out “The release of data from different official departments is inconsistent. Some data may be stored in various data formats ranging from PDF, Excel and XML. For those people who are in the field of data analysis, it is hard for them to adopt PDF in order to make a historical prediction model” [9].

As a result, they might find themselves on the horns of a dilemma – they want useful and valuable numerical and contextual data in PDF, but at the same time, the conversion of PDF is not efficient by human data insertion.

3.2.2 Web Scraping

Web scraping is applied in the stage of data collection. It provides an accurate and rapid information extraction from the Internet. Generally, web scraping is relied on the implementation of Hypertext Transfer Protocol (HTTP). It acts like “user” to browser the website and capture the image of web content. Figure 8 shows the simple workflow of web scraping.

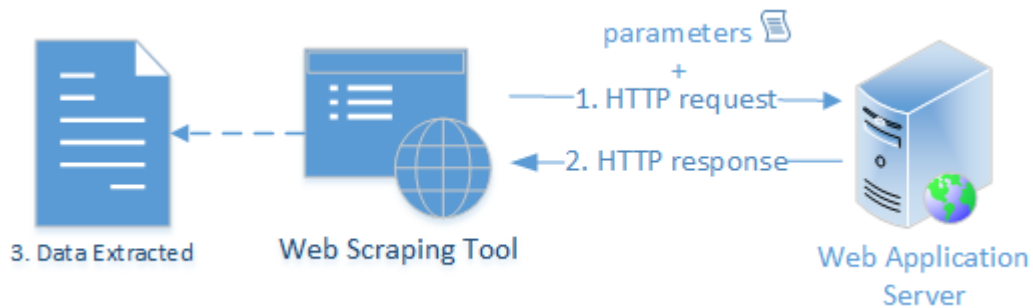


Figure 8 Web Scraping Illustration

Initially, the web scraping tool submits the HTTP request to the web server. The web server will then generate corresponding HTTP response and send back to the client. Also, different content in server pages or other dynamic websites can be captured by passing different parameters of either GET or POST method. Those data collected from web scraping is the raw content of HTML pages, which may include html tags, presentation and scripting. As a consequence, the raw data is required to perform data preprocessing before data analysis and we can use HTML parser to filter out the noise.

3.3 Data Preprocessing

Data preprocessing is an important step to improve the data quality. The removal of irrelevant, redundant and noise can increase the accuracy of data prediction.

3.3.1 Attribute Selection

Attribute selection is a process commonly used in machine learning, where in a subset of the attributes available from the data are selected for application of a learning algorithm. Good attribute selection is able to generalize the model and enhance the performance of prediction. A generalized model can reduce the chance of overfitting since the irrelevant and redundant input features may affect the accuracy of data analysis.

3.3.2 Attribute Creation

Attribute creation refers to the creation of new attributes that can capture more important information than the original attributes in the data.

3.3.3 Data Validation

The data validation includes two parts, which are missing value checking and searching on duplicate tuples. The missing value is defined as blank string, empty string and null value. In Weka, it provides two options on handling missing value. The first one is to replace all the missing value while the other one is to remove all tuples containing missing value. For duplicate tuples case, the redundant tuple will be dropped out from the dataset.

4. Data Modeling

4.1 Transactional Data

JSOUP is a Java library for working with real-world HTML, which provides a very convenient API for extracting and manipulating data, using the best of DOM, CSS and jquery-like methods [10]. First of all, it is needed to find out the servlet or server page responsible for generating the data. Then, CSS selector syntax is applied to select the data. The part of web scraping is consolidated as a standalone program to extract data since it requires much time to capture the transaction records.

The zip file of program can be downloaded from the link:

<http://gproject.appspot.com/rest/download/program>

Unzipping the downloaded file, it should have a text file called README.txt, which provides the introduction and instruction of the program.

```

/*****\
  README.txt for Web Scraping Program

  1.Introduction
  This program is used to scrape the html content
  from the response. It provides a more efficient
  way to capture the data from a specific URL.
  For this program,
  https://www.housingauthority.gov.hk/ha/dweb/en/secMktTrans.do
  The tabular view of transaction records of HOS courts/ TPS estates
  in html table is stored in CSV format named as result.csv

  2.Author
  Lam, Chi Ho

  3.Program Guideline
  First, make sure that you have installed Java
  Next, handle the configuration by editing config.properties file
  Then, execute the "execute.cmd" and wait for the program finishing
  The result.csv will be generated.

  4.Credit and Acknowledgements
  Support by JSOUP library

  5.Change Log
  2016/11/05      Version 1 Initial Version Release

/*****/

```

Figure 9 README.txt in web scraping program

The information of the transaction records is listed in the below table.

Table 2 Attribute of HOS court / TPS estate transaction records

Attribute Name	Description
Year	year of signing date of the provisional agreement for Sale and Purchase
Month	month of signing date
Region	region of flat (Hong Kong, Kowloon, New Territories and Islands)
Location	location of flat
Court / Estate Name	court or estate name of flat
Transaction Price	transaction price (converted into '10,000)
Saleable Floor Area	saleable floor area in sq. ft. only
Floor	floor (H: high floor, M: middle floor, L: low floor)
Transaction Price per Area	extended attribute from calculation of price and area
Discount Rate	premium discount rate
Purchase Method	agency (A) or self-negotiation (S)

Note: The time range of dataset from Jan,2012 to Dec,2016 is chosed.

It is assumed that the housing price is equal to the summation of housing intrinsic value and the floating marketing price by external economic factors. With such assumption of the housing price, the price in transaction record is labelled as class label. In other words, it means that the other attributes expect price are the independent variables to the class attribute – price. Hence, the attribute that is not fit the prediction model should be removed.

As the attributes of transaction price and saleable floor area are available in the transaction records, the derived attribute of “transaction price per area” is removed. The age of building is added to enrich the information.

4.2 Time Series Data

Data	Description
Heng Seng Index (HSI)	The stock and property markets has long been investigated for their correlation. Hang Seng Index is a market capitalisation-weighted index of the consitituent stocks, which is widely used as a benchmark of the Hong Kong stock market.
Consumer Price Index (CPI)	It measures the change in the price level of consumer goods and service purchased in the market calculated by statistical mehod. It can show the consumer sentiment at that time.
Best Lending Rate	With low lending rate, it encourages the will of potential buyer to purchase property. It is because low lending rate can reduce the interest pay to the bank.
Unemployment Rate	The unemployment rate describe how many people is under the period of unemployment over the whole population.
Number of domestic households	It implies the demand of housing.
Centa-City Index (CCI)	It reflects the secondary private residential property price with certain popular real estate as benchmark.

4.2.1 Dependent and Independent Variables

The dependent variable is Centa-City Index while the other variables are independent variables to construct the model of regression.

4.2.2 Multiple Linear Regression

Let the dependent variable be y ,
the independent variables be $\{x\}$,
the unknown parameters be $\{\beta\}$,
the error term be ε .

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

Multiple Linear Regression is a power technique to estimate the relationship among variables. Weka can help to construct the equation of the multiple linear equation.

4.2.3 Interpreting the correlation coefficient r

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y}$$

For correlation coefficient r , the range of r is bounded between 1 and -1. The variables are positively correlated if the value of r is close to 1 while the variables are negatively correlated if the value of r is close to -1.

5. Model Evaluation

5.1 K-fold cross-validation

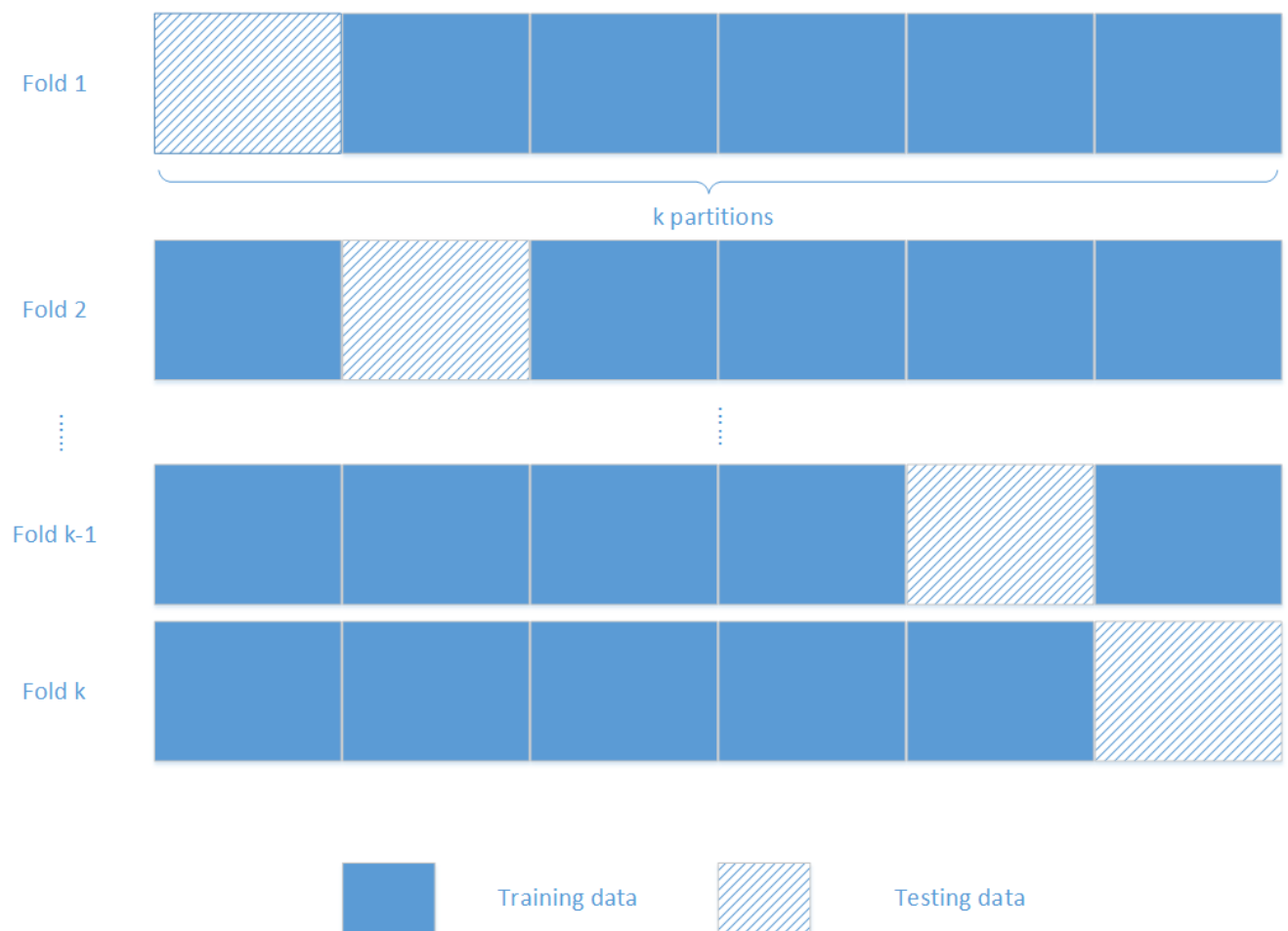


Figure 10 Illustration of K-fold cross-validation

To compare the performance of different classification algorithms, k-fold cross-validation is applied to examine the accuracy that the prediction model gives a correct result. By reference to Figure 10, the sample data is first randomly partitioned into k subsets with equal size. The process will repeat k times. Each time a single data subset is selected for testing the model whereas the remaining k-1 data subsets are used as training data.

5.1.1 Reason of selecting 10-fold cross-validation

As the sample size of transaction records is greater than 20,000, each one tenth of data subset is sufficient for validation. In addition, 10-fold cross-validation gives a more reliable result than hold-out method in spite of the fact that the cross validation takes more time.

5.1.2 Error Rate

The trained prediction model will produce the prediction result of testing data and the corresponding error rate is the result of its incorrect classified instances divided by the number of instances in the sample dataset.

Let E be the error rate of prediction.

$$E_i = \frac{\text{number of incorrect classified instances}}{\text{total number of instances}}, \text{ where } i \in [1, k]$$

After k trials of result validation, the error rate of k -fold cross-validation is calculated. The lower error rate, the better is the evaluation of that classification algorithm.

$$E = \frac{1}{k} \sum_{i=1}^k E_i$$

5.1.3 Confusion Matrix

Apart from the perspective of error rate calculated from cross-validation, Weka provides a detailed illustration of the performance by means of confusion matrix and ROC graph.

Table 3 Confusion Matrix

		Prediction	
		True	False
Actual	True	True Positive (TP)	False Negative (FN)
	False	False Positive (FP)	True Negative (TN)

In case of a binary classifier system, the confusion matrix is a 2-by-2 matrix containing the number of true positives, true negatives, false positives and false negatives.

```

=== Confusion Matrix ===
      a    b  <-- classified as
14882  2151 |    a = Lower class
 1838  9780 |    b = High class

```

Figure 11 Confusion Matrix in Weka User Interface

The above figure shows the sample output of confusion matrix in Weka. The value of false negative is 2151, which means those high class instances are incorrectly classified as the lower class. And the value of false positive is 1838, which means those lower class instances are incorrectly marked as the high class. The total incorrect labeled instances are the summation of FN and FP. Hence, the error rate 'E' of classification is:

$$E = \frac{FN + FP}{TP + TN + FP + FN}$$

The confusion matrix is able to derive not only the error rate of the prediction but also sensitivity. Sensitivity refers to the proportion of lower classes (true cases) that are correctly identified.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

5.1.4 Receiver Operating Characteristic (ROC) Curves

Unlike confusion matrix, ROC curve specifies the evaluation of binary classifier system. ROC curve is a two-dimensional graph resulted by the true positive rate (TPR) against the false positive rate (FPR). Generally, a good predictor should have a smaller FPR than TPR. Below figure shows the graphic interpretation of ROC curve.

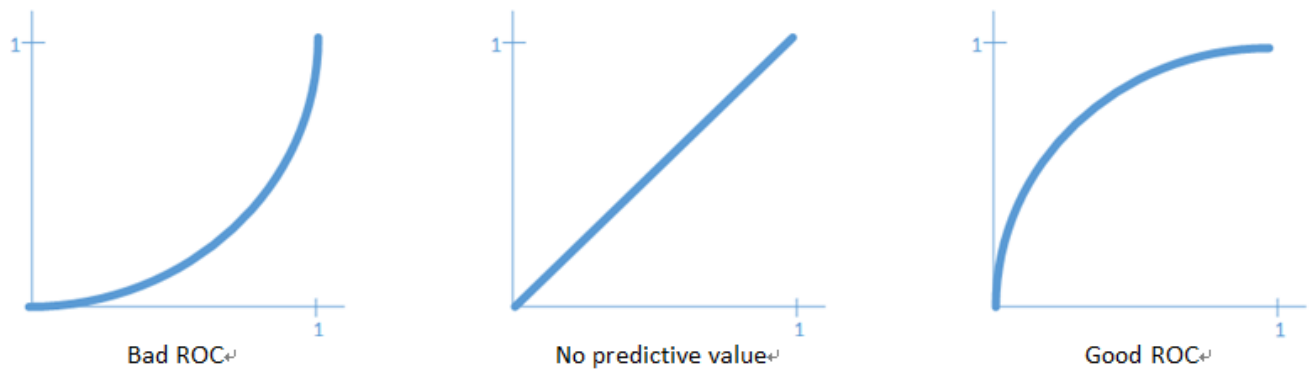


Figure 12 Interpretation of ROC curve

The graphic interpretation only gives a rough estimation of performance. Therefore, the area under ROC curve, which is also called AUC, will be introduced to give a precise numeric output. The best value of AUC is 1.

To simply the evaluation, we use the error rate, sensitivity, specificity and AUC as a combined key performance index of confusion matrix for comparison among three classification algorithms.

Table 4 Performance between various Weka Algorithms

	Naïve Bayes	IBk	J48
Time taken to build model	0.06s	0.01s	1.3s
Error rate	13.9227%	9.7798%	5.937%
Sensitivity (Lower High)	87.4% 84.2%	92.3% 87.2%	94.4% 93.6%
AUC	0.944	0.952	0.965

The above table shows the evaluation result of transactional data model with binary classes, which are high class (greater or equal to \$2,500,000) and lower class (smaller than \$2,500,000). As a whole, it is clear that all three classifiers get a good prediction

6. System Design

6.1 Use Case Diagram

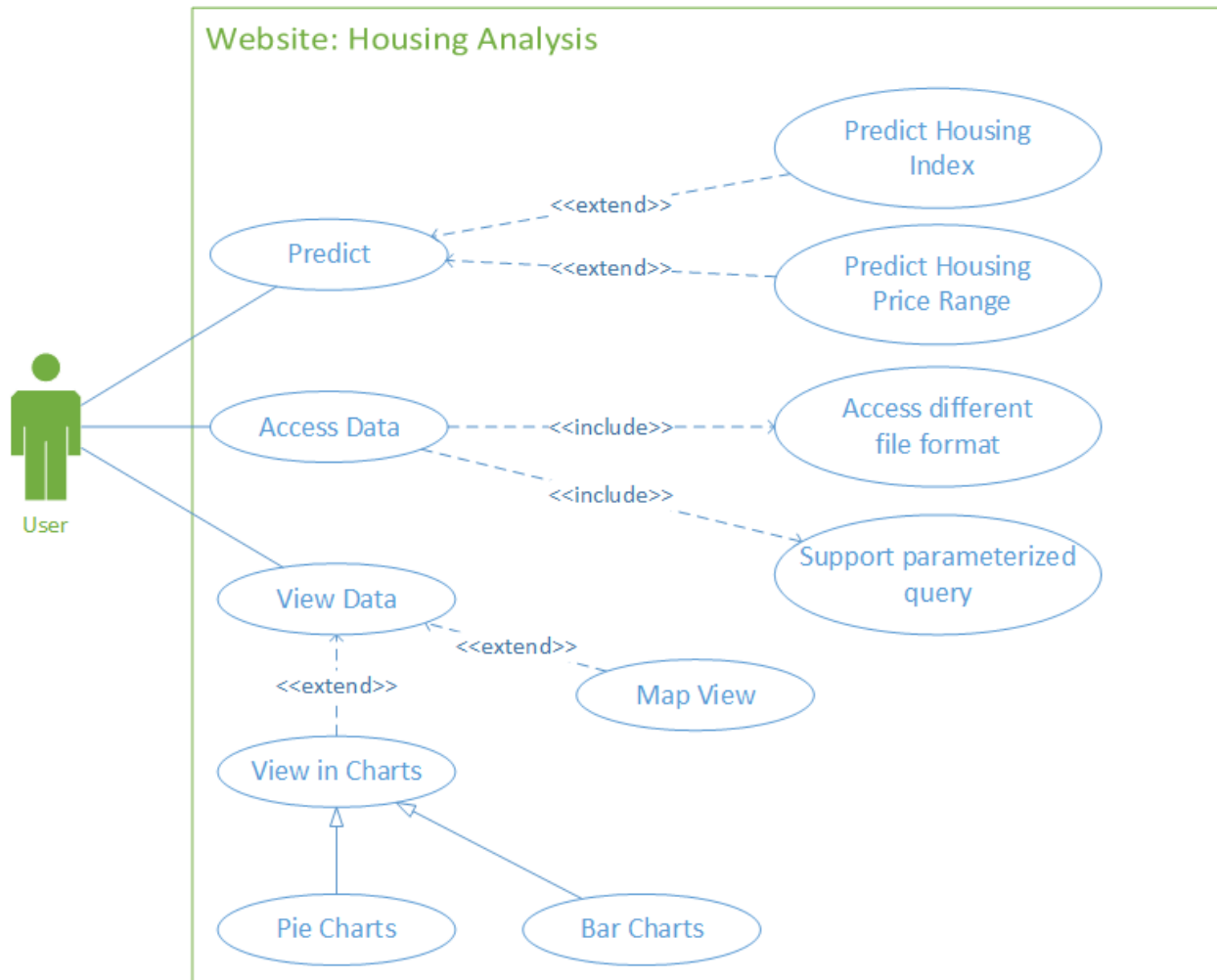


Figure 14 Use Case Diagram

The above diagram shows the interaction of user with the web application system. Different use cases indicate the relationship between user and the system.

6.2 Use Case Description

Use Case	Description	Actor
Predict	User can make the prediction on housing valuation and housing index by submitting the form to the server.	User
Access Data	Using the framework of REST, the web application serves as an endpoint for the access to data with various data formats like XML and JSON. Also, user can make use of data API to perform certain parameterized query to filter out the desired information via HTTP GET method.	User
View Data	User can view the data in a more user-friendly way – an interactive map view, beautiful charts to show the data. User can have a more clean understanding on the dataset.	User

6.3 Sequence Diagram

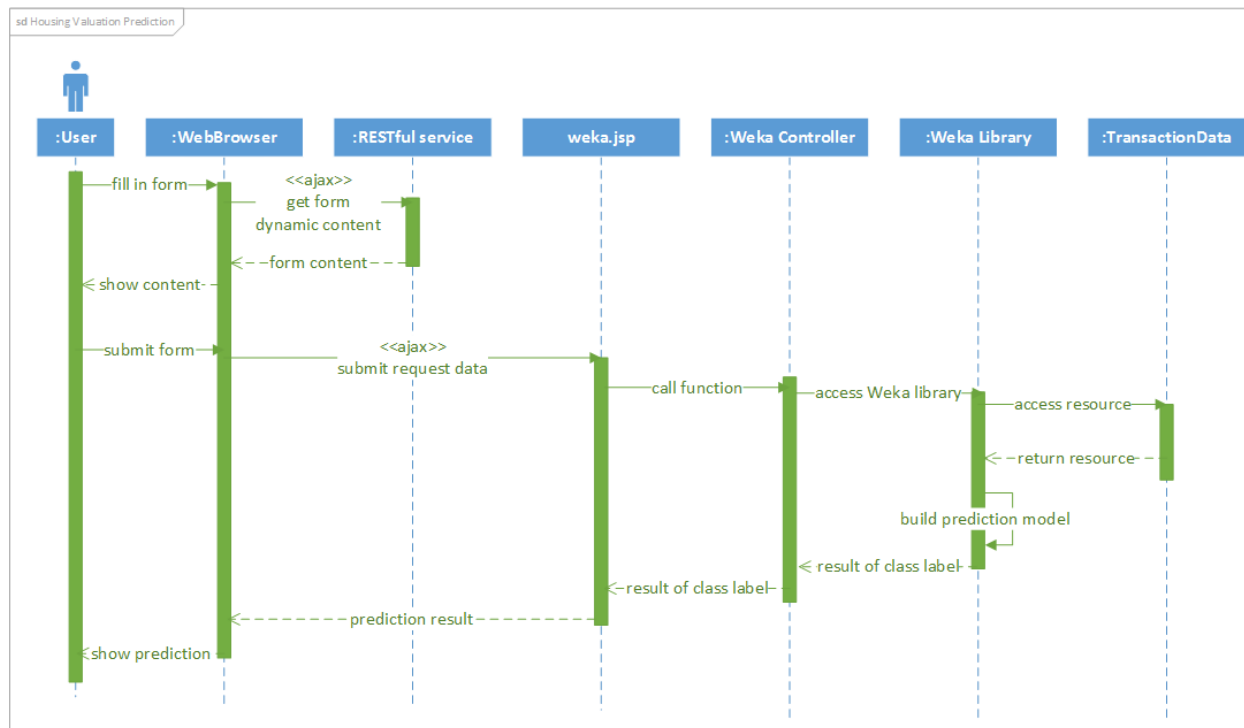


Figure 15 Sequence Diagram - Data Prediction

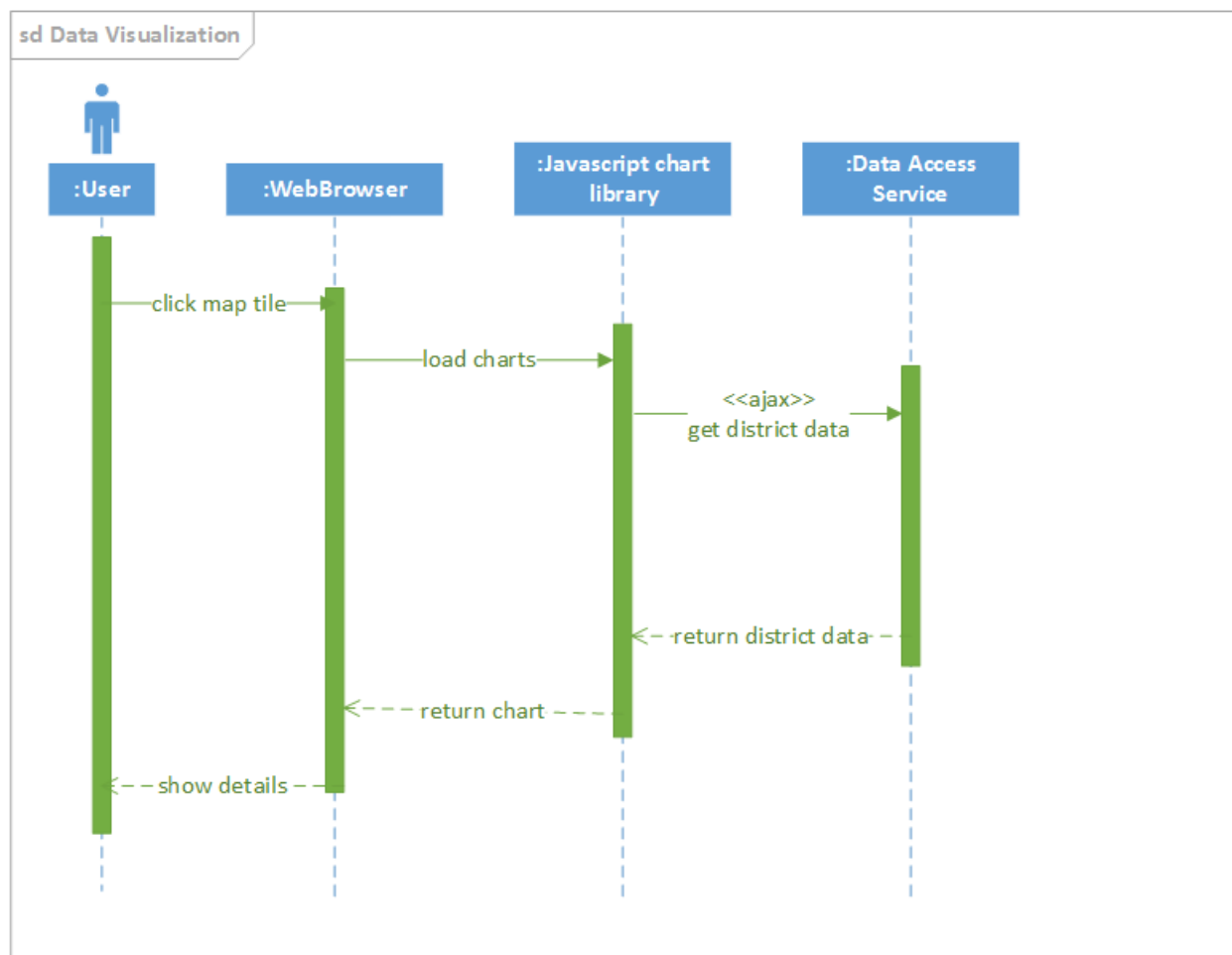


Figure 16 Sequence Diagram - Data Visualization

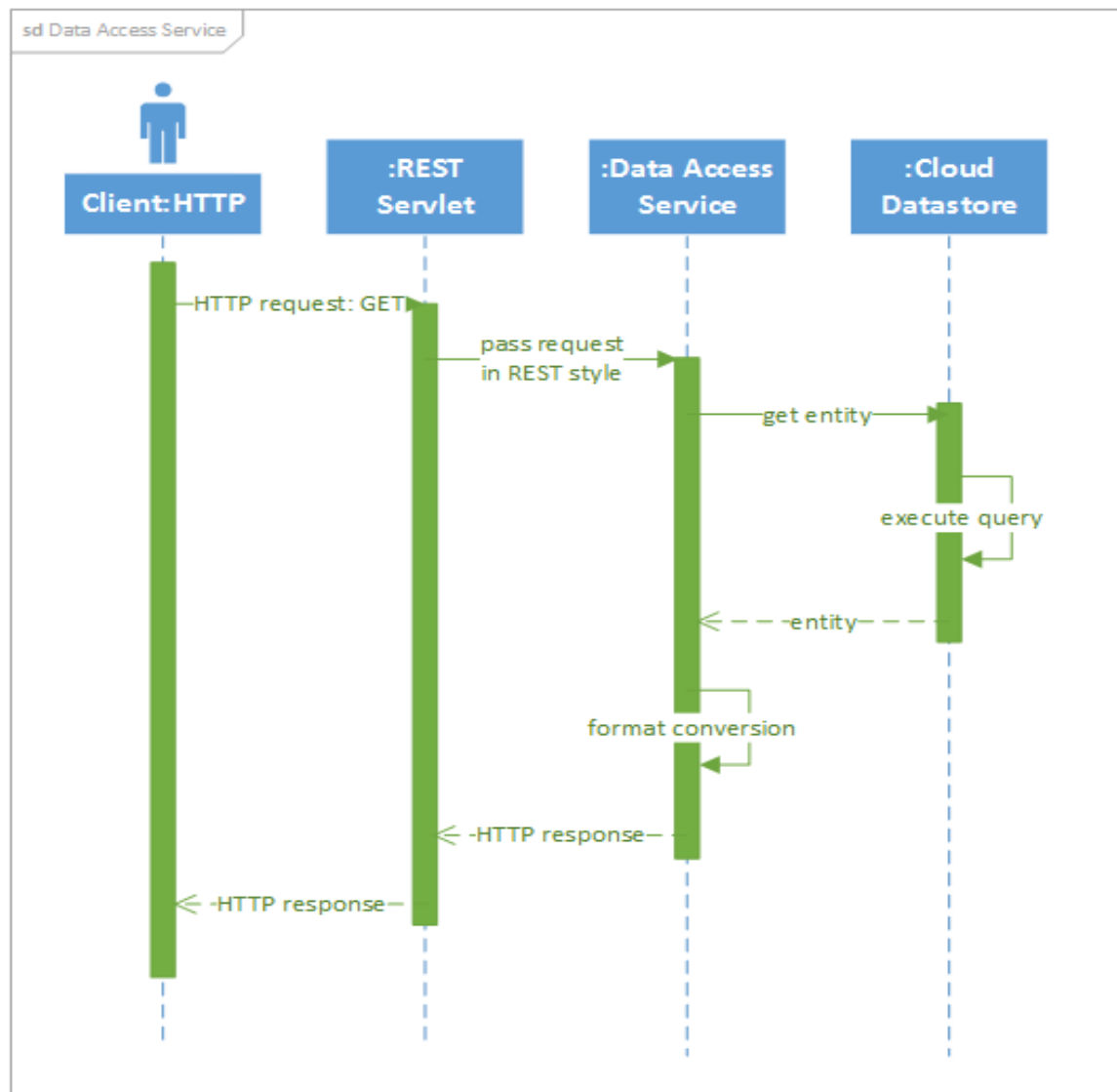


Figure 17 Sequence Diagram of Data Access

6.4 Website Layout Prototype

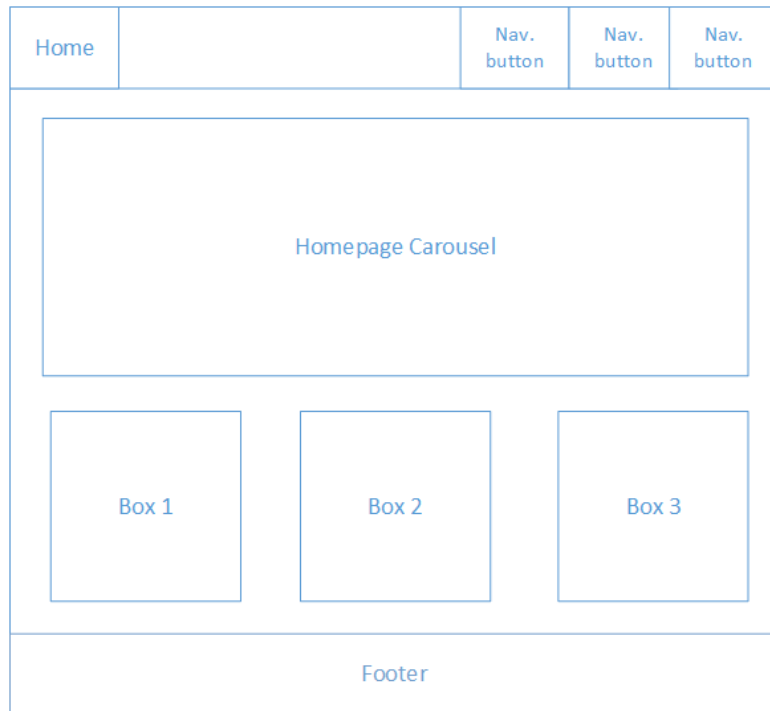


Figure 18 Homepage Prototype

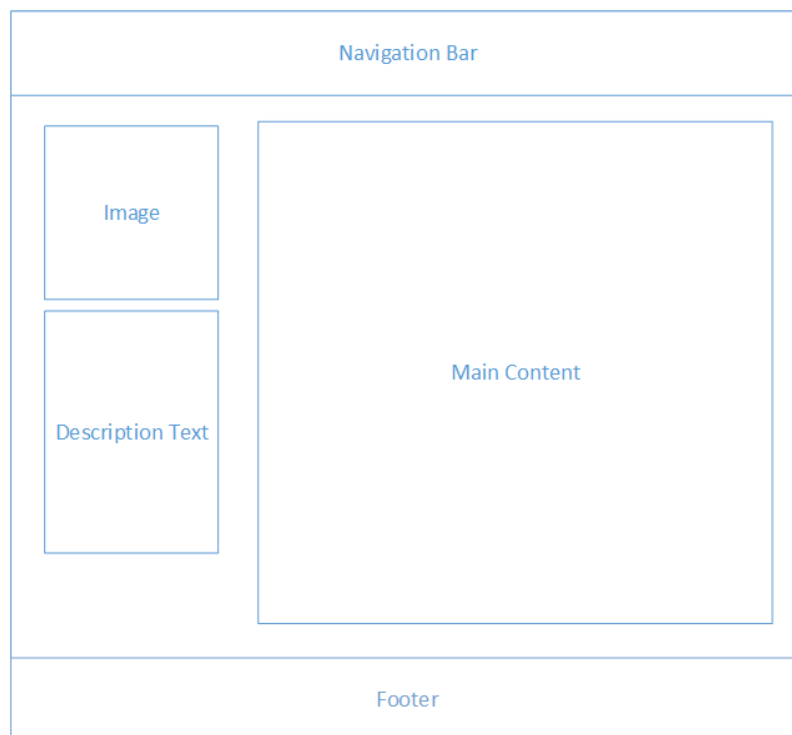


Figure 19 Subpage Prototype

The website adopts Bootstrap CSS framework, which is a responsive web design framework. The above diagram shows the view in desktop devices, for other smaller device like mobile phone, it may have some difference.

7. System Implementation

7.1 Overview of System

The system is a web-based application written in Java, which allows users to access the service and resources through HTTP connection. The server-side of the system is mainly responsible to the business logic and data analysis of housing datasets while the front-end framework is using the Bootstrap CSS framework and JavaScript to construct the presentation and layout of website.



Figure 20 Overview of System

7.2 Core Programming Language



Java is selected as the main programming language of the system because of its open source nature and platform independent characteristics. Java can run in different operating system so it is flexible and convenient. Despite the fact that the main purpose of Java is not designated for statistical computing, some external library written in Java enrich the power of data analysis.

7.3 Selected Platform and Tools



7.3.1 Waikato Environment for Knowledge Analysis (Weka)

Weka is well-known machine learning software written in Java. It offers both graphical user interface and Java API library for easy access to the algorithms for predictive modeling and data analysis. In the table, it shows the corresponding Weka class for selected algorithms. In previous Section 2.3, the concept of each data mining algorithms are generally explained. The unpruned configuration of decision tree set to

be false, which can improve the speed of building decision tree model. The number of k in kNN sets to be 3 - the most optimistic number of k by trying several trial tests.

Table 5 Weka Selected Algorithms for Classification

	C4.5 Tree	Decision structure	Naïve Classifier	Bayes based	k-Nearest Neighbors
Working principle	Tree-like classifier		Conditional probability model		Instance-based learning
Weka implementation	J48 class		NaiveBayes class		IBk class
Class configuration	Unpruned: false		None		Euclidean distance Value of k: 3

7.3.2 Google App Engine (GAE)

Google App Engine is a cloud service for developing web applications and it currently support multiple programming languages – Python, Java, Go and PHP. App Engine enables the flexibility to extend the application scale.

However, there are some limitations in GAE like read-only access to the filesystem on App Engine.



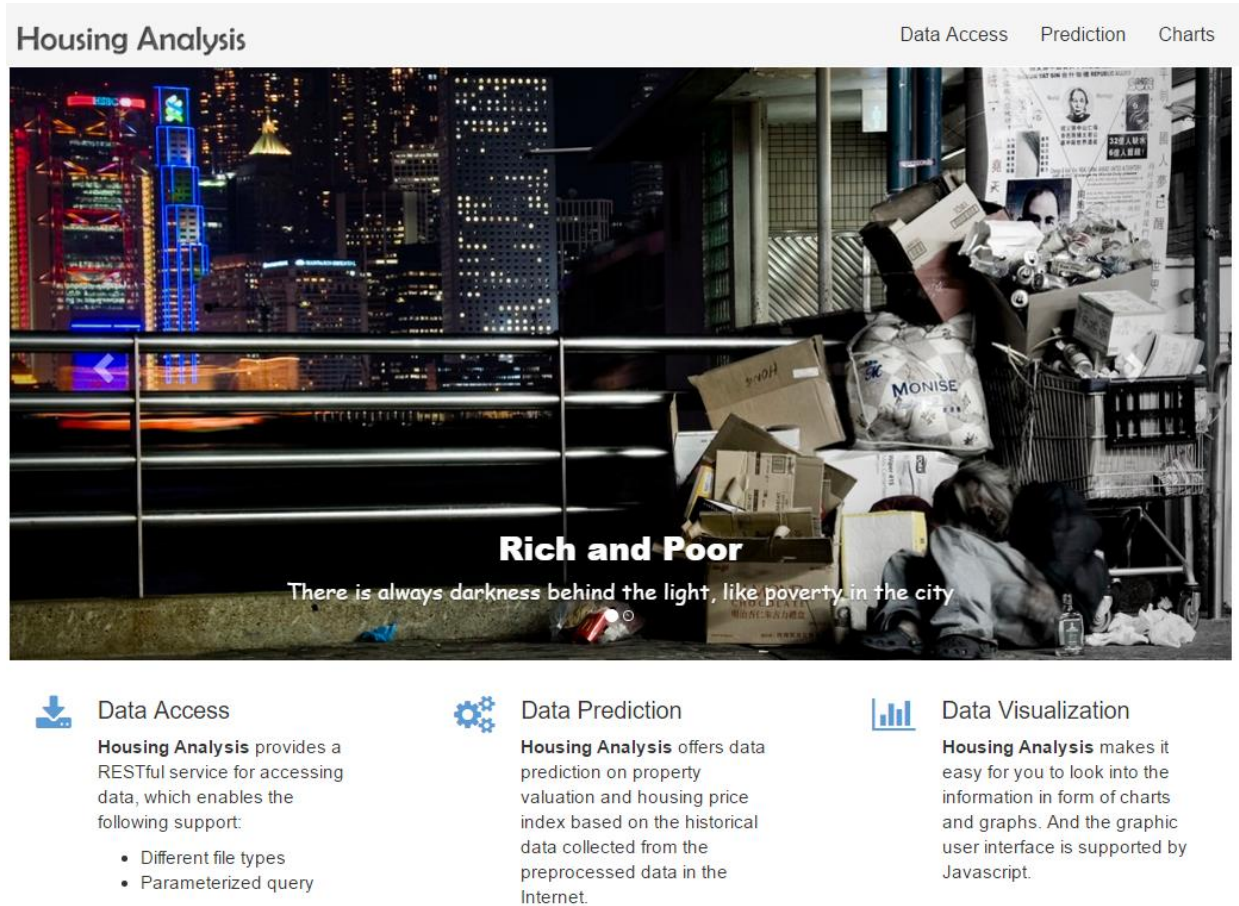
7.3.3 Google Cloud Datastore

Google Cloud Datastore is a good companion with Google App Engine. It is a NoSQL document database built for automatic scaling, high performance, and ease of application development [11].

7.4 Web Application

Please refer to the link of web application here <http://gprojectapi.appspot.com/>.

7.4.1 Homepage



Housing Analysis Data Access Prediction Charts

Rich and Poor
There is always darkness behind the light, like poverty in the city

Data Access
Housing Analysis provides a RESTful service for accessing data, which enables the following support:

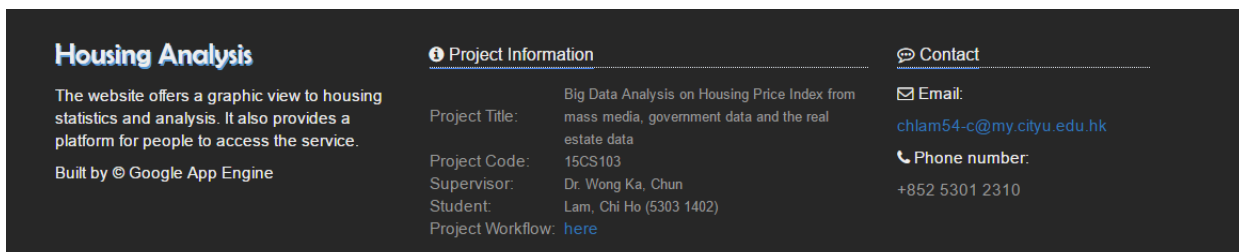
- Different file types
- Parameterized query

Data Prediction
Housing Analysis offers data prediction on property valuation and housing price index based on the historical data collected from the preprocessed data in the Internet.

Data Visualization
Housing Analysis makes it easy for you to look into the information in form of charts and graphs. And the graphic user interface is supported by Javascript.

Figure 21 Screen Capture - Homepage

7.4.2 Footer



Housing Analysis
The website offers a graphic view to housing statistics and analysis. It also provides a platform for people to access the service.
Built by © Google App Engine


Project Information
Project Title: Big Data Analysis on Housing Price Index from mass media, government data and the real estate data
Project Code: 15CS103
Supervisor: Dr. Wong Ka, Chun
Student: Lam, Chi Ho (5303 1402)
Project Workflow: [here](#)

Contact
Email: chlam54-c@my.cityu.edu.hk
Phone number: +852 5301 2310

Figure 22 Screen Capture - Footer

Housing Analysis

Data Access
Prediction
Charts



Data Access

Using the framework of REST, *Data Access* provides the access to data with various data formats like XML and JSON. Also, the data API allows you to perform certain parameterized query to filter out the desired information via HTTP GET method.


Available Datasets


Consolidated Time-series Data of Housing Price Factors
🔼


🕒 Latest Update on Dec, 2015


The statistics generally relate to the consolidated time-series data related to housing price from various data sources. It includes Hang Seng Index(HSI), Consumer Price Index(CPI), best lending rate, unemployment rate, number of domestic households and Cenci-Central Index(CCI). The time range of the dataset is from Jan, 1994 to Nov, 2015.

Data Resources

 CSV

 JSON

 XML



GET: http://gprojectapi.appspot.com/rest/hfi/<data_format>?param1=value¶m2=value

Supported parameter	Format of value
data_format	{"csv", "json", "xml"}
year	numeric string
month	numeric string
download	{"Y", "N"}

Figure 23 Screen Capture - Data Access

In section 3.2.1, it mentions that the government open data is difficult to extract its information from the file format of PDF and Excel. Owing to the fact that data extraction from PDF and Excel is hard, the website is designed to build up the service to provide an endpoint accessing data, which allows the usage for both users and system. The style of representation state transfer (REST) suits for the service of data access. The RESTful service of the website mainly provides two major functionalities. The first one serves for people to access or download different file format besides PDF and Excel formats. The output file formats like XML and JSON are self-describing structure and they can be serialized into object without heavy loading and processing. The second one is to support the parameterized query to filter out the desired information. The framework of RESTful data access service is Java API for RESTful Web Service (JAX-RS) and Java Architecture for XML Binding (JAXB) handles the conversion to XML and JSON data formats. According to the practice of RESTful style, the HTTP GET method is always used to read a resource or collection. The below diagram shows the flow of data access service.

3. Merge the content into HTTP response

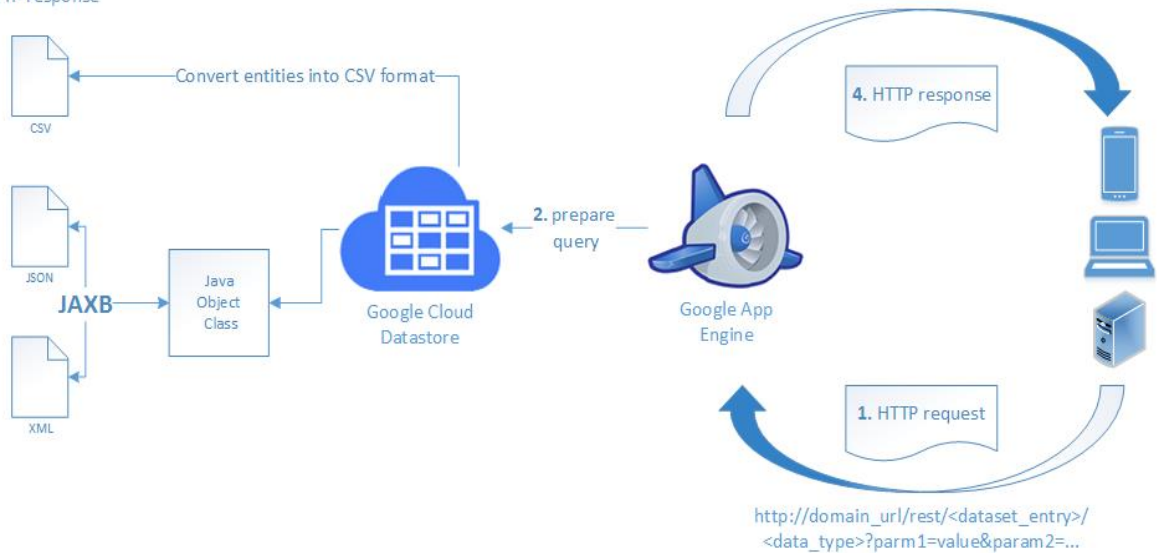


Figure 24 RESTful data access service flow diagram

URL Design for Restful Web Service

URL	<a href="http://gprojectapi.appspot.com/rest/<dataset_key>/<data_format>">http://gprojectapi.appspot.com/rest/<dataset_key>/<data_format>	
Method	GET	
Parameter in path	dataset_key	{“hf”, “ncf”}
	data_format	{“csv”, “xml”, “json”}
Querystring	Refer to the guide in http://gprojectapi.appspot.com/data_access	
Return	200 OK	

Example

URL	http://gprojectapi.appspot.com/rest/ncf/json? year=2012&district=Eastern
Method	GET
Return	200 OK
	Output: <pre> {"NCF":[{"count":"0","district":"Eastern","quarter":"3","type":"PRH","year":"2012"}, {"count":"0","district":"Eastern","quarter":"4","type":"PRH","year":"2012"}, {"count":"0","district":"Eastern","quarter":"3","type":"HOS","year":"2012"}, {"count":"0","district":"Eastern","quarter":"4","type":"HOS","year":"2012"}, {"count":"108","district":"Eastern","quarter":"3","type":"PH","year":"2012"}, {"count":"0","district":"Eastern","quarter":"4","type":"PH","year":"2012"}, {"count":"0","district":"Eastern","quarter":"1","type":"PRH","year":"2012"}, {"count":"0","district":"Eastern","quarter":"2","type":"PRH","year":"2012"}, {"count":"0","district":"Eastern","quarter":"1","type":"HOS","year":"2012"}, {"count":"0","district":"Eastern","quarter":"2","type":"HOS","year":"2012"}, {"count":"0","district":"Eastern","quarter":"1","type":"PH","year":"2012"}, {"count":"0","district":"Eastern","quarter":"2","type":"PH","year":"2012"}]}</pre>


7.4.4 Data Prediction

Housing Analysis

Data Access

Prediction

Charts



Data Prediction

There are two sets of prediction form in *Data Prediction*, which are prediction on housing valuation built by Weka C4.5 decision tree classifier and prediction on Centa-City Index built by linear regression model.

Property Valuation on HOS court /TPS estate

Year

Month

Region

Please select region

District

Please select district

Location

Please select location

Estate

Please select estate

Area

Please enter saleable area in sq. ft. unit

Floor

Please enter floor

Rate(%)

Please enter discount rate

Channel

Please select channel

Age

Please enter property age

Confirm

Reset

Figure 25 Screen Capture - Data Prediction

There are two sets of prediction in Data Prediction, which are prediction on housing valuation built by Weka C4.5 decision tree classifier and prediction on Centa-City Index built by linear regression model. The data prediction is achieved by the form in web application. User can fill in the input text field and select the option. And then click “Confirm” button to view the prediction result.

Region

Please select region

Please select an item in the list

Figure 26 Screen Capture of responsive input field

Those two prediction form have responsive input form fields, which assist the user to fill in the form. If the input form field is filled with incorrect value, it will show red in color like Figure 26. Otherwise, it will show green in color to indicate the filled information is correct.

7.4.5 Data Visualization

Data Visualization is designated to provide an alternative view of data rather than table format to present data. The choropleth map can show the density of statistic in geographic plane. Furthermore, the user is able to interact with the choropleth map with hover and click action. When the user use the mouse hover the map tile, it shows the corresponding statistics.

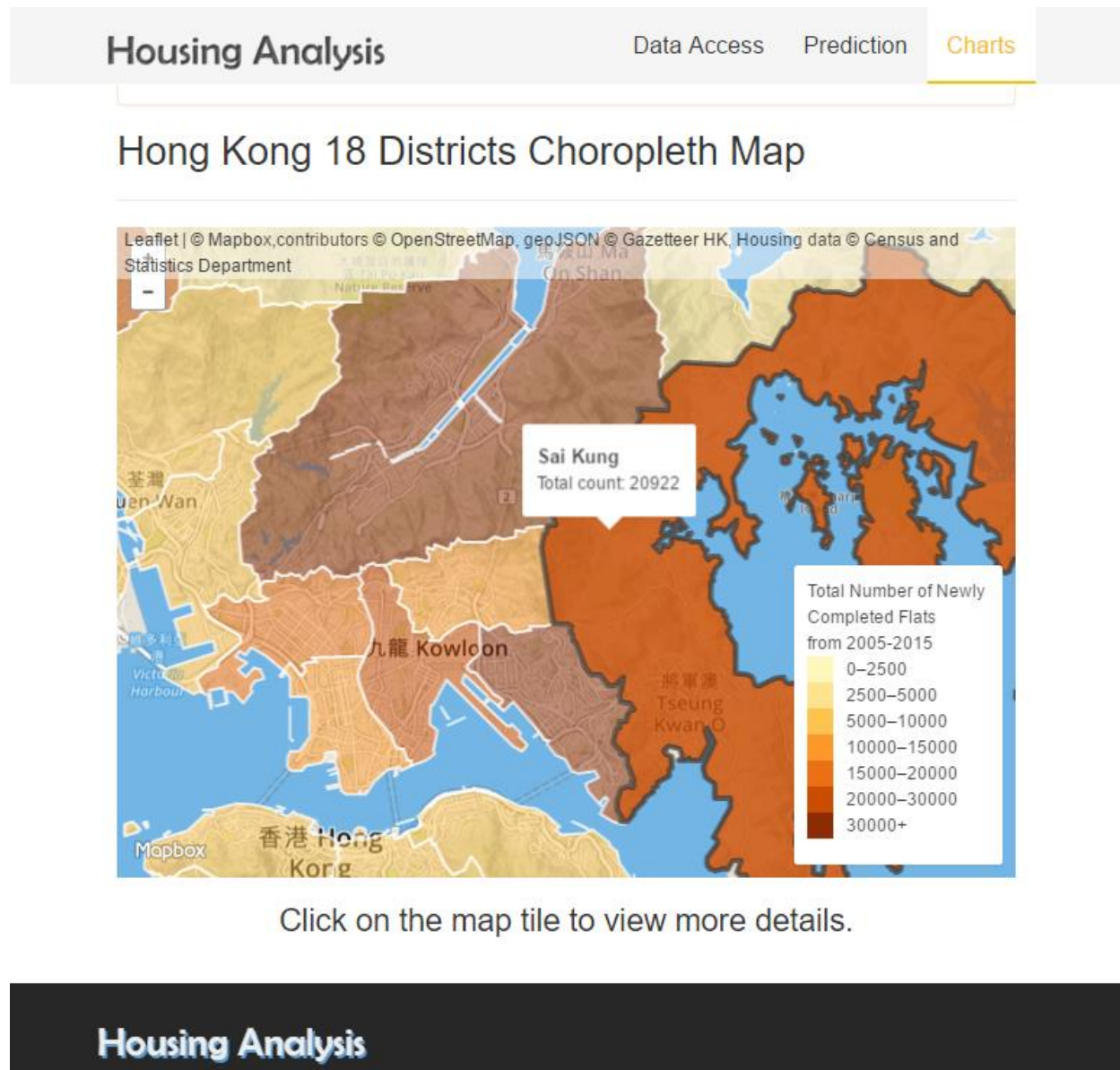


Figure 27 Screen Capture - Data Visualization

When the user click on that map tile to view more details, it will then show the below figure.

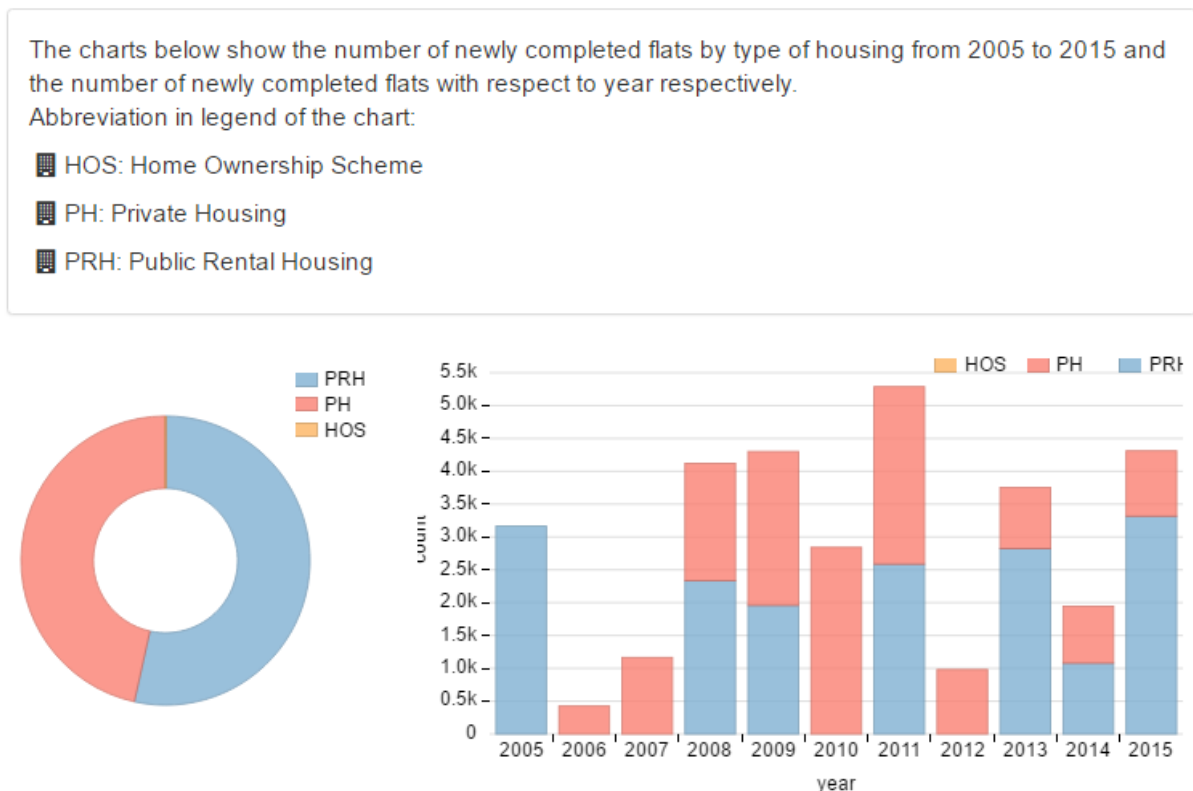


Figure 28 Detail View on statistics

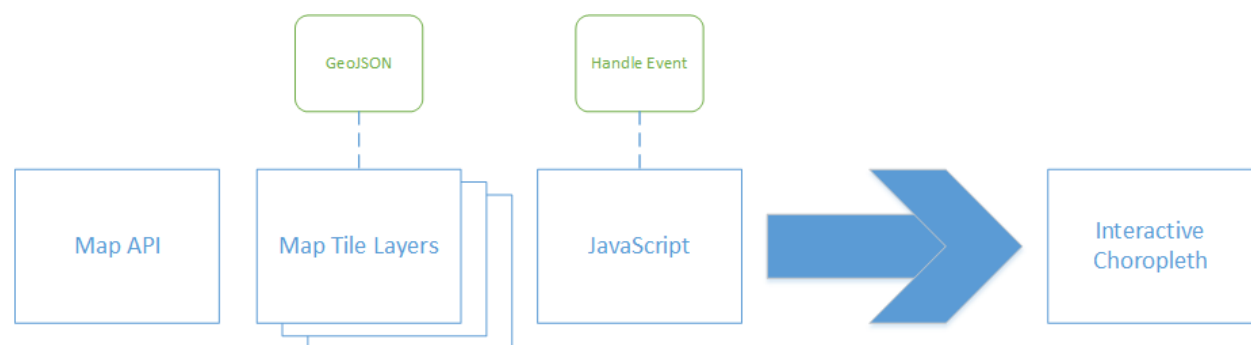


Figure 29 Principle of Interactive Choropleth Map

The idea is inspired by Gazetter HK. Owing to the license of Google Map, the add-on map tile is not allowed. The choropleth map is supported by Mapbox, which is a open source map platform for developers. The map tile layers is formed by GeoJSON file, which is formed and edited by the online tool with the link <http://geojson.io/>. Combining the map api and map tile layers, it forms the basic choropleth map in static view. Finally, jQuery in JavaScript is applied to control the user's event.

8. Project Review and Conclusion

When comes to 2016, the housing price in Hong Kong falls from the peak in the first quarter of 2016. Honest Speaking, the initial motivation of this project is related to the extreme high housing price in Hong Kong. At that time, I as young generation felt quite hopeless and upset when looking on the soaring increase in price level of real estate market. I decided to choose analyzing housing problem as my topic of final year project topic. After studying the situation, I realize that it is more important to stabilize the housing price, which means the government should cool down property market when its price increase rapidly. The time and strength of housing control measure is the key. Therefore, effective and accurate prediction could help.

As the project has limited time to work on, there is a room for the system to extend and improve its functionality. Data insertion to database is one of the example. Prediction model is built by historical data. While the program does not have the automation of data update. Therefore, the reliability of data prediction on recent time will be improved if the historical data can be accumulated.

The assumption checking for multiple linear regression model. Actually, there should be some hypothesis made on the regression like checking on non-existence of multicollinearity, normality and homoscedasticity.

Appendix I – Monthly Progress Log

Month	Log
September 2015	(a) Background research and study (b) Define the problem and scope (c) Documentation of project plan
October 2015	(a) Data collection <ul style="list-style-type: none"> Manual method Web scraping (b) Data Source <ul style="list-style-type: none"> Government open data Real estate statistics Mass media (c) Design on system architecture
November 2015	(a) Data pre-processing (b) Data cleaning <ul style="list-style-type: none"> Remove noise Enhance consistency
December 2015	(a) Data study (b) Data analysis
January 2016	(a) Data analysis on time series data <ul style="list-style-type: none"> Correlation between housing price index and hang seng index (HSI) (b) Data analysis from housing transaction data
February, 2016	(a) Evaluation the performance of the classifiers using k-folds cross validation (b) Debugging and Testing (c) Coding Part (d) Integration with Weka library

	<p>(e) Website layout</p> <p>(f) Website functionality</p>
March, 2016	<p>(a) Documentation:</p> <ul style="list-style-type: none"> • Documentation of final report • User case diagram of website and other diagram that helps to illustration the flow and functionality of website <p>(b) Programming:</p> <ul style="list-style-type: none"> • Build up the RESTful service for data access • Continue testing on the web platform

References

- [1] Yvonne Liu (2014) South China Morning Post “Hong Kong ranks world's No 1 for 'most unaffordable' housing” retrieved Sept 15, 2015, from <http://www.scmp.com/property/hong-kong-china/article/1410730/hong-kong-ranks-worlds-no-1-most-unaffordable-housing?page=all>
- [2] Hong Kong Housing Authority (2015) Housing in Figures 2015 report
- [3] Teo Nicolais (2014-15) Website – “How to Use Real Estate Trends to Predict the Next Housing Bubble” retrieved Nov 15, 2015, from <http://www.dce.harvard.edu/professional/blog/how-use-real-estate-trends-predict-next-housing-bubble>
- [4] Hing Lin Chan, Kai Yin Woo (2011) “Studying the dynamic relationships between residential property prices, stock prices, and GDP in Hong Kong” Hong Kong Shue Yan University
- [5] Mehmed Kantardzic (2011, p.184) IEEE Press “Data Mining Concepts, Models, Methods, and Algorithms”
- [6] Everitt, B. S., Landau, S., Leese, M. and Stahl, D. (2011) Miscellaneous Clustering Methods, in Cluster Analysis, 5th Edition, John Wiley & Sons, Ltd, Chichester, UK.
- [7] Sintef (2013) ScienceDaily “Big Data, for better or worse: 90% of world's data generated over last two years” retrieved Oct 28, 2015, from <http://www.sciencedaily.com/releases/2013/05/130522085217.htm>
- [8] FAQ (n.d.) In Data.gov.hk retrieved Feb 28, 2016 from <https://data.gov.hk/en/faq>
- [9] [Data-HK]. (2014, March 16). RTHK - How Open Data changes society? [Video File] retrieved from Feb 28, 2016
- [10] Jsoup: Java HTML Parser (2015) Website retrieved Nov 20, 2015 from <http://jsoup.org>

[11] Google Cloud Platform (March 25, 2016) Website – “What is Google Cloud Datastore?” retrieved Feb 28, 2016 from <https://cloud.google.com/datastore/docs/concepts/overview>