Knn Algorithm used in

Real Estate price estimation

|  |
| --- |
|  |
|  |
| Table of Contents  [Two-senetence summary 1](#_Toc15144024)  [summary 1](#_Toc15144025)  [The problem 1](#_Toc15144026)  [The solution 1](#_Toc15144027)  [The customer experience 2](#_Toc15144028)  [The Business impact 2](#_Toc15144029) | |  |  |

# Two-senetence summary

Based on the known information, finding the best prediction of the price.

Not the highest or lowest, but the nearest.

Examples:

If we have information about a list of real estate’s prices and their features, then we can use KNN method to make some

predictions based on what information we have.

The ultimate goal is to find the nearest price, but not the highest or lowest price.

# summary

Short descriptions:

1) Feature Comparison

2) Deploy KNN algorithm

3) Price estimation

# The problem

1. How to compare the feature between two lists and how to deal with different data types

2. In deploying KNN algorithm, what type of distance should we use?

3. After getting the result, how can we evaluate our results and model?

# The solution

In the following chapter, I will use jupyter notebook to solve the problem.

1. Data Preprocessing

After loading the dataset into the notebook, first we need to check the basic info of this dataset.

A screenshot of text

Description automatically generated

We can find that we have 2450 observations and 27 columns in this dataset. Different columns have different data types, and some columns may contain null value. Therefore, first we need to transform some numerical data types into categorical type, like PROPERTY, ZIPCODE.

Also, we transformed some features to make obvious distinction among them. For example, we square the BEDROOMS so as to increase the separation between not many bedrooms(1bedrooms) and many bedrooms(4). In this way, the difference between them from 3(4-1) increase to 15(16-1). We also transformed some features like squaring the BEDS\_SQUARE, taking the square root of SQFT\_SQRT and LOT\_SIZE.

After that, we need to select, what features should we include that can be the most relevant to predict the prices. Here, we choose 9 features, including square\_feet, lot\_size, BEDS,BATHS and so on. After we choose the features we want to analyze, we create a new variable to store the feature and relevant data and drop the null-values in this new dataset.

Apart from that, in order to effectively calculate the distance between different points, we need to convert this DataFrame into 2d array. And because some features may have much larger number than other(e.g. Square\_feet and BEDS), in order to consider all the features at the same level, we need to normalize this to exclude some influence of extreme values.

Now, data preprocessing is done.

2. Perform KNN algorithm

The full name of KNN is K-Nearest-Neighbors. Here, k is a number, ‘nearest’ meaning the closest or smallest, neighbors are like sorted value. Then, we may infer the KNN is to find the number of K smallest values from a sorted data. The ultimate goal of KNN is to do classification and make prediction.

While performing KNN algorithm, there are a lot of ways to calculate the distance between two points. Here, in order to choose the best method to do the calculation, we try to take into account all the possible ways to do this.

2.1 Euclidean Distance

When we talking about distance calculation, the first method will come into our mind is this method. This is the typical way to calculate the distance. This method takes all the data points into consideration, including outliers. The formula of this method is as follow:

2.2 Cosine Distance

Cosine Similarity is a metric used to determine how two sets are similar. This method is more advantageous because it ignores the size of each sets and more accurate results. Cosine Distance is more used in text and document analysis, for example, this can be used to calculate the frequency words and find the most frequent word. The formula is as follow:

2.3 Jaccard Distance

The Jaccard Index is a measurement to compare the similarity between two sets to check what they have in common and what is distinct. In my opinion, it is used useful to do some corrections when doing some experiments or manufacturing. Because we have some different features to compare between the new one and the standard one. And we may find what their differences and can better improve the production procedure or the experiments. The formula is as follow:

2.4 Chebychev Distance

This method compares each variable’s difference and choose the maximum as the estimation. In my opinion, because I have took the Project Management and Operation Research, it is more used to estimate the maximum time, cost and route when preparing business plan. For example, if I can know the maximum time and cost that I may use for this project, then I can expect the worst situation and think of a way to deal with this situation to save the time and cost and make more profit. The formula is as follow：

2.5 Mahalanobis Distance

Mahalanobis Distance is used for calculating the distance between two data points in a multivariate space. The advantage of using this method is that it takes into account of the covariance and helps to measure the similarity.

However, as far as I am concerned, this method is not useful in KNN algorithm. Because we can just get each test data’s distance separately, and may not help us to use the minimal distance to find the estimated price.

The formula is as follow:

X is the vector of the observation.

M is the vector of mean values of independent variables.

D is the inverse covariance matrix of independent variables.

After thinking of all the ways to calculate the distance, we can see which methods perform best and choose the best K.

3. Methods Evaluation

Here, we will take the average prices of KNN as prediction. And in order to get quicker results, we will take k from 1 to 5 to compare the results.

There are plenty of ways to evaluate the performance of this method. We can calculate the MAE, MSE, MAPE. The algorithm behind these methods is the same, that is to calculate the difference between the actual value and prediction. But these methods stand in different levels. Here, we are going to use RSS to evaluate the models.

RSS means residual sum of squares. Residual means the difference between the actual value and prediction. The formula of RSS is:

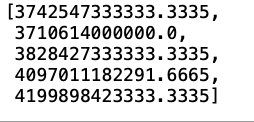
If we use the Euclidean, we can get a result like this:

A screenshot of a cell phone

Description automatically generated

The smallest RSS is 3.344911e+14 when k is 5.

If we use Cosine method to calculate the distance, the result is like this:



The smallest RSS is 3710614000000 when k=2.

If we use Jaccard to calculate the distance, the smallest RSS is 3.422016e+12 when k is 5.

If we use Manhattan to calculate the distance,

A screenshot of a cell phone

Description automatically generated

The smallest RSS is 3.280121e+14 when k=5.

If we use Chebyshev to calculate the distance,

A screenshot of a cell phone

Description automatically generated

The smallest RSS is 3.372855e+14 when k is 5.

Therefore, to sum up, we can get a table like this:

|  |  |  |
| --- | --- | --- |
| Method | RSS | K |
| Euclidean | 3.344911e+14 | 5 |
| Cosine | 3.710614e+14 | 5 |
| Jaccard | 3.422016e+12 | 5 |
| Manhattan | 3.280121e+14 | 5 |
| Chebyshev | 3.372855e+14 | 5 |

The smallest RSS is using the Manhattan method.

However, we may notice that, in the above results, we may all see a downward trend using k from 1 to 5. However, if k is increasing, the RSS will continue to decrease or not?

In order to see this trend, we need to consider to increase the range of k.

If we use k from 1 to 15, the Euclidean’s result is like this:

A screenshot of a cell phone

Description automatically generated

There may exists a downward trend but also a upward trend. Here, the smallest RSS is 2.978470e+14 when k is 14.

Therefore, it is not wise to use a small range of k-value. In order to make more precise estimation, we need to do more experiments using larger ranges of k.

# The customer experience

The ultimate to get the more accurate prediction of the price using some features. Therefore, for the customer, we can provide this service to some real estate developers or consulting firms, so they can make better strategy of it. For example, for developers, when they choose an area and they need to decide whether to invest money or not. Then, they can use this method, they need to name the basic features and their plans, then we can get the estimation of this price. It will help them make a decision.

And for consulting firms, they can use this method, to provide more accurate price to their customers. The reputation of this consulting firm will increase because of their precise estimation.

# The Business impact

Using this method to estimate the price of real estate, it will be more convenient and accurate. However, the database still needs to be updated frequently to get more accurate estimations.