Our Team Name: 取名字真累

Our Team Members: 陈文帅、于鹏、叶恺、罗绍博、王程灏、茅宇润

Abstract:

Part 1: What's the background and the meaning to predict the price

Part 2: How can we do?

Part 3: Our program's problems & improvement

Part One: The Importance of House Location

- **background**

The housing price in Beijing and Shanghai is always high, but recently when I viewed a website, I found it even more incredible. A house of 50 square meters is about 2-2.5 million in Xicheng District in Beijing, what happened?

As we all know, Beijing is crowded and also still economically going strong, as is Shanghai. The tech boom has certainly brought even more of the rich and wealthy demographic to Beijing, and Xicheng is right in the center of the city. The schools there are good, so you can avoid all the hassle and expense of getting your child into a good school by just declaring the Xicheng house your residence and living in a bigger place elsewhere in the city.

The real estate market in Shanghai has been thriving recently, with low interest rates and limited inventory creating an ideal environment for sellers in the city. Buyers, on the other hand, often face escalating real estate prices, bidding wars and prolonged search periods as they enter an increasingly competitive market. This is particularly true in Shanghai and Beijing, which are home to the nation's two most expensive housing markets. So how exactly are real estate prices determined? We've uncovered three of the most important factors affecting the housing market and that will almost certainly impact the price of houses per square meter.

Economists encapsulate “location” in something called “hedonic pricing” – for most homes, this translates to some key factors that impact your life and your lifestyle: Real estate prices are heavily influenced by location and neighborhood.

Quality of local schools is frequently the single most important factor for buyers with children of school-going age; Proximity to local employment opportunities is a very high priority for most employment-age buyers; Proximity to social, shopping and recreational centers is valued most by younger buyers but plays an important role in pricing for all homebuyers;
These factors are not independent of one another – e.g., many parents want to drop the kids off and pick them up at school as part of a reasonable commute to and from work. These three preferences – proximity to school, work and entertainment/shopping — are a trinity that make for immensely valuable property. Generally, getting a home which has one of the three attributes won’t blow the proverbial roof off the price per square meter. If you’re getting two out of three, one should expect stiff competition and commensurate prices. To get all three, one might need a small war chest to finance your home-buying exploits.

Therefore, we find some most obvious factors such as school district information, subway information, number of parks to figure out how these factors influence the price. In order to quantify the information, we decide to collect the distances of these houses to the nearest subway station, to the nearest school and to the nearest park. We may choose a most suitable algorithm from KNeighbors Regressor, Linear Regressor, Adaboost Regressor, Decision Tree Regressor and Random Forest Regressor. By using the suitable algorithm, we will get the relationship between housing price and some factors so that we can predict some new residential area’s price.

Part Two: How to predict the price

- **Solutions**

The process of programming and the result

Before we started to program, we happened to find that the there is hardly any data for us to find, thus we just have collected 200 lines of data that can come in handy. However, the statistics we find are fairly complicated, it caused us long time to train the model and get the value of MSE. In this case, we created a file called 'price. Csv'. Someone may doubt that our data is not that precise and thus increase the errors in our prediction. But we carefully checked our data and tried to ensure that the limitation of data won't have too much impact on the result.

After we collected the data, we were going to find the best algorithm in training the machine and predicting the price of housing per square meter. Here are the algorithms we want to try:

1. KNeighbors Regressor
2. Linear Regressior
3. Decision Tree Regressor
4. Random Forest Regressor

And we used the metric MSE (Mean Square Error) for measuring the performance of a regression model.

The following are the detailed procedures by which the python program works:
First, the python program reads the new price data using [pandas]. As we can see in the chart, the data is divided into two parts. One is 3 features of different housing, including the distance of the nearest subway station, the distance of the nearest school and park.

Then, we prepared feature matrix X and label vector y, and use holdout validation to split X and y into two parts: training (80%) and testing (20%).

As the value X and Y are prepared, it is time to train a model .We used the algorithms (KNeighbors Regressor, Linear Regressior, Decision Tree Regressor, Random Forest Regressor) in turn (Not in a loop because it will take the computer a long time to run the code).

After we trained the model, we were going to evaluate the regression model on testing part. We ran the program to compute the MSE of each algorithm for further comparing. It seemed to work quite well. The following chart shows what the program has got:

Algorithm	MSE
KNeighbors Regressor	5.891485
Linear Regressor	0.986493
DecisionTree Regressor	1.383725
RandomForest Regressor	0.655586

Accordingly, we can draw a conclusion that the algorithm of KNeighbors Regressor doesn’t fit our case, and the Random Forest Regressor performs the best.

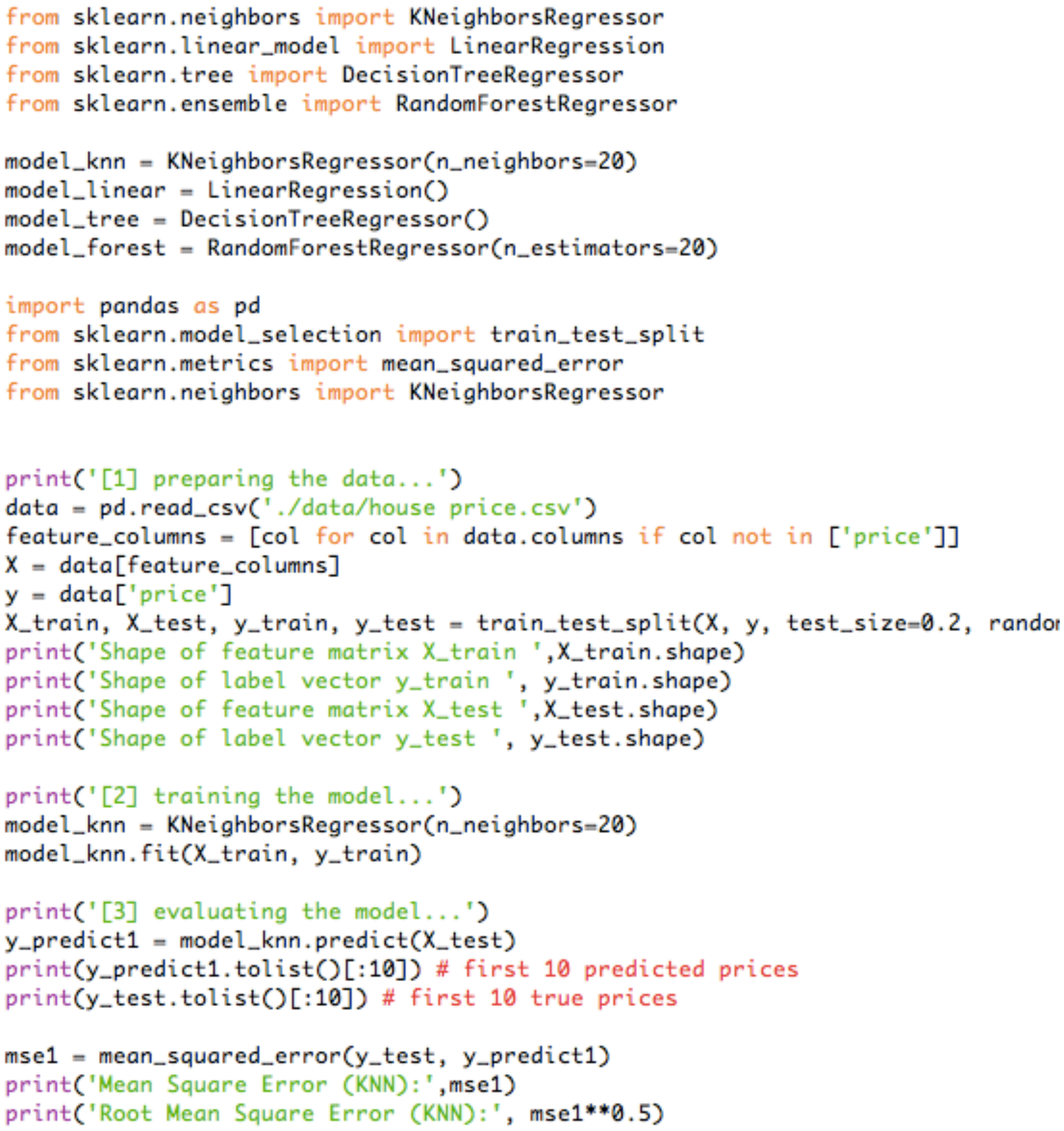
Since we have found the best algorithm to predict the price of the housing, now it’s time to use the program and predict the housing prices which we don’t have in the market. When we apply the algorithm we have chosen, as soon as the user input the three kinds of distance information into a file called “predict price.csv” .

And then users can run the program, they can learn about the prices of houses about to sell.Our goal is to have a relationship between housing relative features and its price. So at last it print the results.

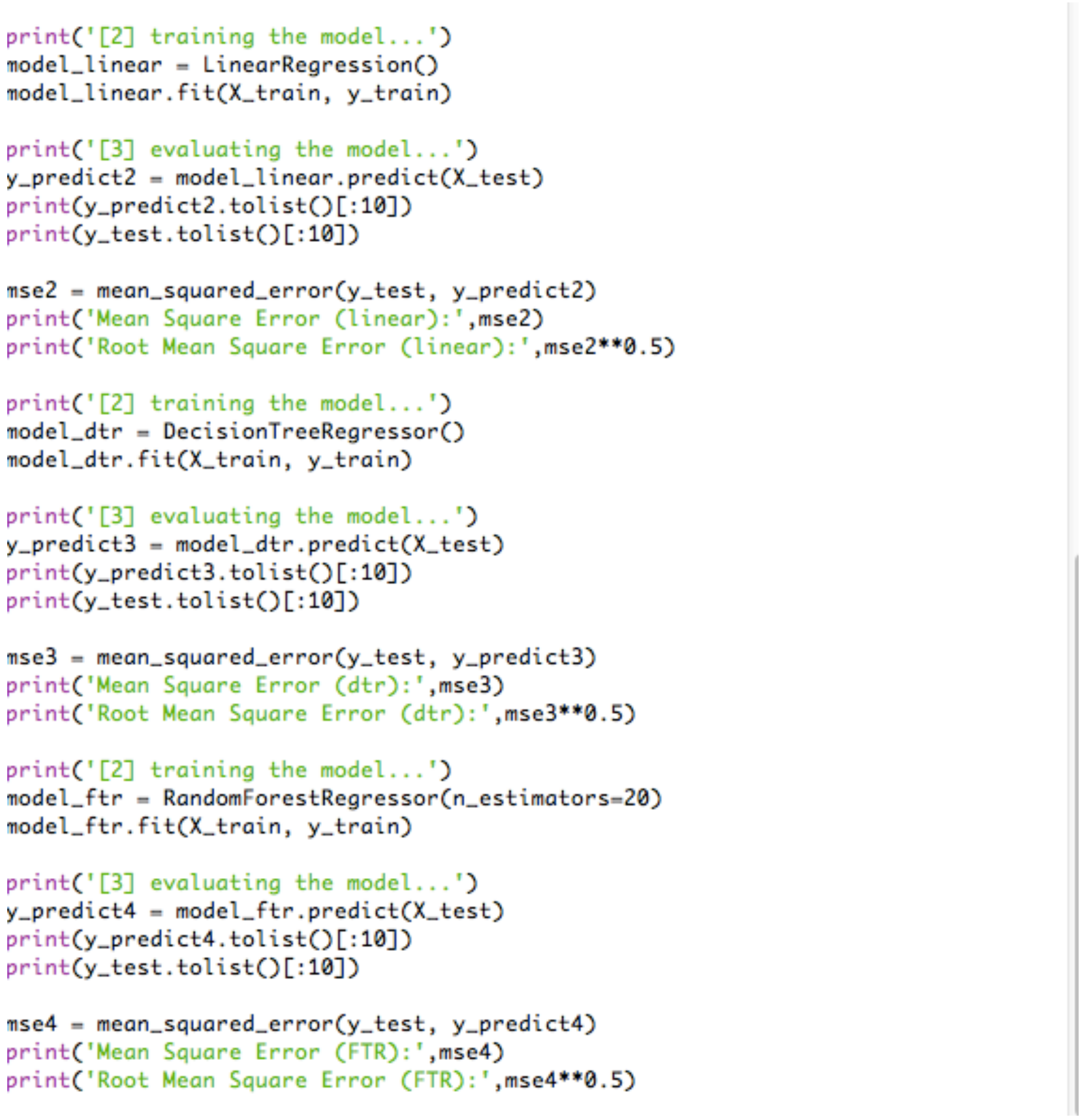
That’s the whole process of our programming, and we have also shown our thinking process in it.

This is our code：

Stage1：

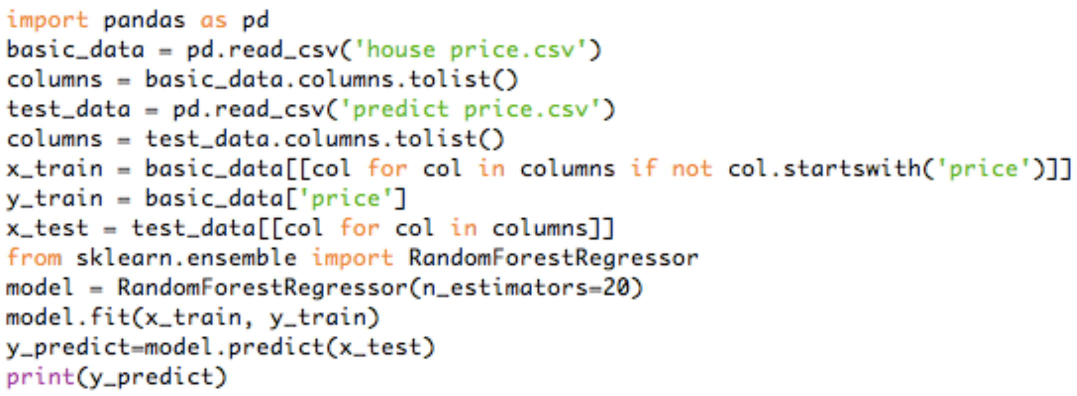


```
from sklearn.neighbors import KNeighborsRegressor
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor

model_knn = KNeighborsRegressor(n_neighbors=20)
model_linear = LinearRegression()
model_tree = DecisionTreeRegressor()
model_forest = RandomForestRegressor(n_estimators=20)

import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from sklearn.neighbors import KNeighborsRegressor


print('[1] preparing the data...')
data = pd.read_csv('./data/house price.csv')
feature_columns = [col for col in data.columns if col not in ['price']]
X = data[feature_columns]
y = data['price']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, rando
print('Shape of feature matrix X_train ',X_train.shape)
print('Shape of label vector y_train ', y_train.shape)
print('Shape of feature matrix X_test ',X_test.shape)
print('Shape of label vector y_test ', y_test.shape)

print('[2] training the model...')
model_knn = KNeighborsRegressor(n_neighbors=20)
model_knn.fit(X_train, y_train)

print('[3] evaluating the model...')
y_predict1 = model_knn.predict(X_test)
print(y_predict1.tolist()[:10]) # first 10 predicted prices
print(y_test.tolist()[:10]) # first 10 true prices

mse1 = mean_squared_error(y_test, y_predict1)
print('Mean Square Error (KNN):',mse1)
print('Root Mean Square Error (KNN):', mse1**0.5)
```



```
print('[2] training the model...')
model_linear = LinearRegression()
model_linear.fit(X_train, y_train)

print('[3] evaluating the model...')
y_predict2 = model_linear.predict(X_test)
print(y_predict2.tolist()[:10])
print(y_test.tolist()[:10])

mse2 = mean_squared_error(y_test, y_predict2)
print('Mean Square Error (linear):',mse2)
print('Root Mean Square Error (linear):',mse2**0.5)

print('[2] training the model...')
model_dtr = DecisionTreeRegressor()
model_dtr.fit(X_train, y_train)

print('[3] evaluating the model...')
y_predict3 = model_dtr.predict(X_test)
print(y_predict3.tolist()[:10])
print(y_test.tolist()[:10])

mse3 = mean_squared_error(y_test, y_predict3)
print('Mean Square Error (dtr):',mse3)
print('Root Mean Square Error (dtr):',mse3**0.5)

print('[2] training the model...')
model_ftr = RandomForestRegressor(n_estimators=20)
model_ftr.fit(X_train, y_train)

print('[3] evaluating the model...')
y_predict4 = model_ftr.predict(X_test)
print(y_predict4.tolist()[:10])
print(y_test.tolist()[:10])

mse4 = mean_squared_error(y_test, y_predict4)
print('Mean Square Error (FTR):',mse4)
print('Root Mean Square Error (FTR):',mse4**0.5)
```

Stage2 ：



```
import pandas as pd
basic_data = pd.read_csv('house price.csv')
columns = basic_data.columns.tolist()
test_data = pd.read_csv('predict price.csv')
columns = test_data.columns.tolist()
x_train = basic_data[[col for col in columns if not col.startswith('price')]]
y_train = basic_data['price']
x_test = test_data[[col for col in columns]]
from sklearn.ensemble import RandomForestRegressor
model = RandomForestRegressor(n_estimators=20)
model.fit(x_train, y_train)
y_predict=model.predict(x_test)
print(y_predict)
```

Part Three: The Problems We are Faced with

- **Problem**

1. Since house prices vary among numerous regions and our code is not able to reflect the changes in prices due to the region;
2. We compare the distances between houses and the nearest subway lines. However, we can’t compare the differences between these subway lines and the intensive degree of them, thus we have no idea of how these factors affect the house prices. For instance: the house next to the Xuhui subway station is more expensive than the house next to the Dongchuan Road subway station
3. We only consider the relationship between housing prices and the nearest school, but different schools have different effects on house prices, for example: high schools and primary schools near the house affect its price per square meter differently.
4. We do not consider the impact of the differences between different parks on housing prices, for example: the house next to the Xuhui subway station is more expensive than the house next to the Dongchuan Road subway station
5. Sample data collection is very difficult, and our limited sample data has a certain impact on the selection of K value of KNN method.

- **Improvement**

We think what we can improve is about the prediction of the price, because we just let users input the three kinds of information into the “predict price.csv”. We think it is so easy that the users may feel unsatisfied. So we may upgrade the process of inputting the statistics.

Also, the factors we consider can expand to the distance of the city center , etc. But the collection of the distance may be difficult, users should measure the distance on the electric map on their own. So we can expand the factors and try to find some other easier ways to collect statistics.

Write at the end of the report:

Thanks for Bao Yang and assistants for the hard work of the whole term. The introduction course is necessary but a bit difficult for the beginners(: D), but we will do our best to study the knowledge. Anyway, thanks for all your work, attention and sweat!