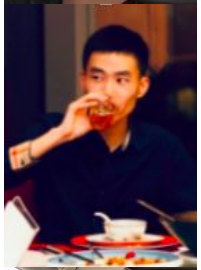
YYDS Group Project

House Prices of Beijing, Shanghai and Guangzhou

Group Member Introduction

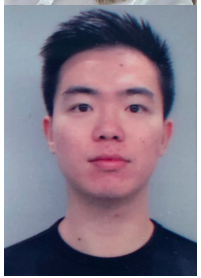
Chunhe Zhang

Xinghui Li



Rong Yong

Shulin Gu



Reason topic was selected

After graduation, we may face the problem of buying a house. So, we are curious about the house price in China, and also the trend for it.

With our analysis, we know what kind of house is cheaper and we can predict the house price of any kind of houses.

Description of the source of data

We selected some data of Beijing House Price from Kaggle

https://www.kaggle.com/ruiqurm/lianjia

All Data in Kaggle is from the following website

https://bj.lianjia.com/chengjiao.

but thats not enough, to make our analysis more complex and meaningful, we did a lot of web scraping and got House Price in Shanghai and Guangzhou as well.

Moreover, we not only considered houses that already been sold, but we also considered houses that are currently on sale.

Questions the team hopes to answer with the data

1. Which city’s houses are more expensive.

2. Which district’s houses are more expensive.

3. Create interactive tables and charts for customers and our audience to select their dreaming house.

4. Predict house price for customer’s dreaming house.

Description of the data exploration phase of the project

1. Discussing where to find the big data, and which data to choose
2. Checking the complexity of the data
3. Seeking more data
4. Data cleaning
5. Create the database

Technique used in this project

1.Python

- Web Scrapping (pandas, numpy, BeautifulSoup, ChromeDriverManager, etc)
- Data Cleaning (sqlalchemy, config, request, etc)
- Machine Learning (sklearn, matplotlib, seaborn, sqlite3, etc)
- Flask

2.JavaScript

- HTML
- Dynamic Tables
- Data Visualization (Charts)
- Mapping with JS and APIs

3.SQL

- SQLite

Machine Learning

- Supervised Learning: Regression
- Outcome(y): Total Price
- Features(X):
  - City
  - District
  - FloorLevel
  - FloorType
  - BuildingType
  - BuildingStructure
  - RenovationCondition
  - Elevator
  - Number of Bedrooms
  - Number of Living rooms
  - Number of Kitchen
  - Number of Bathrooms
- LabelEncoder, StandardScaler
- Split 80% data into training set and 20% data into testing set

Machine Learning

```
LinearRegression Mean Squared Error: 85869.92055431839
LinearRegression R2: 0.38539979917493605
DecisionTreeRegressor Mean Squared Error: 55600.36686514736
DecisionTreeRegressor R2: 0.6020492808113085
RandomForestRegressor Mean Squared Error: 40721.46958394869
RandomForestRegressor R2: 0.7085426046440196
LGBMRegressor Mean Squared Error: 36747.29619343738
LGBMRegressor R2: 0.7369871140619224
SVR Mean Squared Error: 149021.97425272633
SVR R2: -0.06660090881459557
KNeighborsRegressor Mean Squared Error: 55938.968084646935
KNeighborsRegressor R2: 0.5996257968234271
Lasso Mean Squared Error: 85869.90741591265
Lasso R2: 0.38539989321097046
```

The result shows LightGBM as our best model since it has the lowest MSE and the highest R2.

DashBoard

Now let’s have a look at our beautiful, gorgeous, remarkable, splendid, amazing dashboard :)))

We aim to create a dashbaord that can link to different pages.

- We will have home page, about us, team, page information for different cities, mapping, etc.
- We use javascript to make the dashboard
- We plan to add what we learn from the module (searching bar, mapping, graphing)

Thank You

Feel free to ask any questions :)