1. Data source

This data is a forecast of apartment rent competition data, which is from Kaggle.

The apartment rent is predicted by apartment information published on the Internet, such as room layout, quantity, text description, photos, price, etc. The apartments are located in New York City.

The data is from the forecast apartment rent competition: (https://data.cityofnewyork.us/City-Government/DOF-Cooperative-Comparable-Citywide-/myei-c3fa)

2. Analysis purpose

Use the historical data with monthly rent tags to establish different models, realize the prediction of monthly housing rent based on basic housing information, and provide an objective measurement standard for the city's rental market.

3. Analyze the problem

This project will use the training set data as the sample training model, use different models to predict the rent according to the prediction set data, compare the performance of different models according to the result error, and hope to determine the model applicable to the monthly housing rent prediction,

Data Discovery (EDA)

Skewness value: it can be used to measure the asymmetry of probability distribution of random variables.

Kurtosis value: it can be used to measure the steepness of the probability distribution of random variables.

View the rent distribution curve and draw a curve distribution diagram and rent scatter diagram、

图形用户界面, 直方图

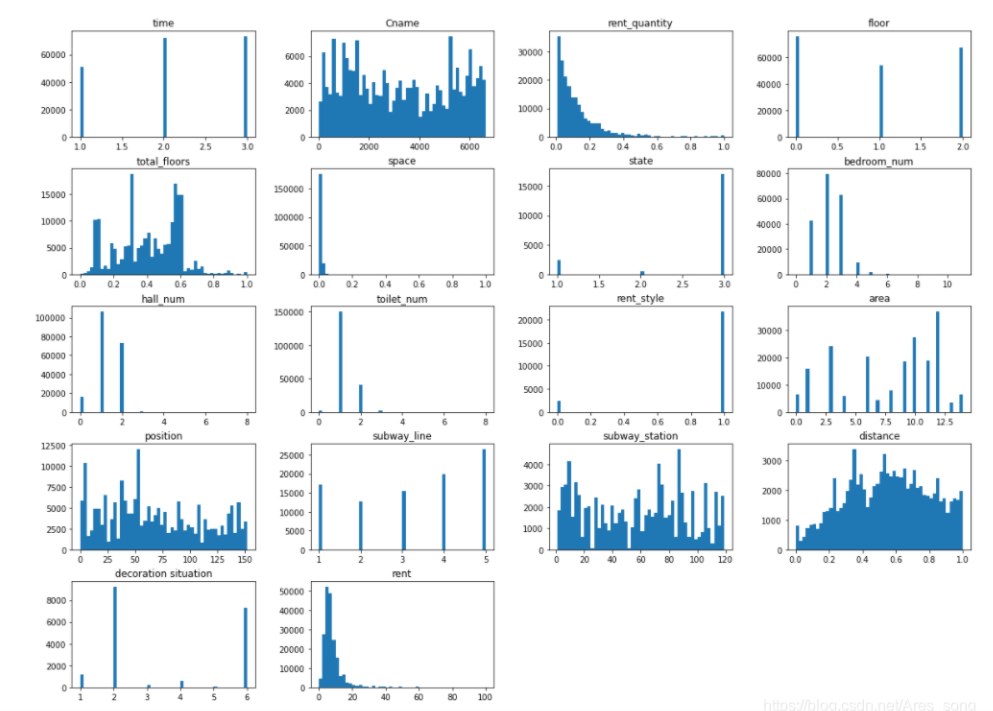
中度可信度描述已自动生成

First, observe the rent data. It can be seen that this data is in line with the normal distribution, but the skewness value is too large. We found a big tail. The "rent" value distribution has obvious skewness, so we will correct it later.

(2) View the data histogram of each property of the apartment

Next, let's take a look at the impact of some important attributes on the results

The histogram is used to show the data distribution. Generally speaking, it refers to which piece of data accounts for a high proportion or number of occurrences, and which piece has a low probability of occurrence. The following figure shows the occurrence of 18 attributes



Continuous variables include: community name (Cname), rent\_quantity, total floors, position, subway\_station, distance, rent, and the rest are discrete variables.

(3) For discrete variables, we use boxplot to represent (box plot)

Discrete variables include time, floor, space, state, bedroom\_num, hall\_num, toilet \_num

Rent\_style, area, subway\_line, decoration.

Box diagram of bedroom number and rent

图片包含 图表

描述已自动生成

Box diagram of living room number and rent

图表, 箱线图

描述已自动生成

Box diagram of area and subway\_line

日历

描述已自动生成

(4) Correlation analysis

Analyze the correlation degree between different factors affecting housing price and housing price, and analyze the correlation degree through the correlation graph.

Qualitative and visual analysis of the correlation between different factors affecting housing price and housing price with bar chart

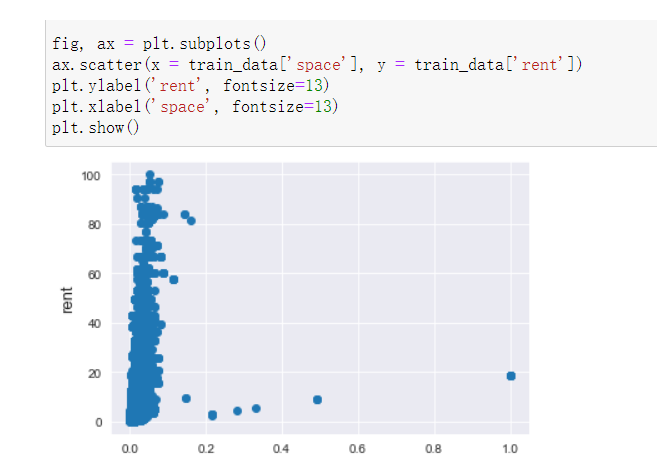
图表, 条形图

描述已自动生成

Data cleaning

Because the data may be incomplete, noisy, random, and have complex data structures, it is necessary to preliminarily sort out the data, clean the incomplete data, make preliminary description and analysis, select variables related to data mining, or change variables.

1. Outlier processing



图表

低可信度描述已自动生成

Deviation correction

图表

描述已自动生成

Normal distribution transformation: the tail behind is too long, and it needs to be corrected (logarithmic transformation of data)

图形用户界面, 应用程序

描述已自动生成

图表, 直方图

描述已自动生成

图形用户界面, 图表

描述已自动生成

For attributes with too large missing data, you can choose to discard them, and for attribute data with small missing data, you can fill them:

(1) Delete the missing data (decoration situation, state, rent\_style)

all\_ data. drop(['decoration situation'], axis=1, inplace=True)

all\_ data. drop(['state'], axis=1, inplace=True)

all\_ data. drop(['rent\_style'], axis=1, inplace=True)

(2) The three columns of "subway line", "subway station" and "distance" represent the subway situation near the house. According to different data types, the distance column filled with "1" represents unlimited distance, and the first two columns filled with "0" represent no subway station and subway line

all\_ data["distance"]=all\_ data["distance"].fillna(1)

all\_ data["subway\_line"]=all\_ data["subway\_line"].fillna(0)

all\_ data["subway\_station"]=all\_ data["subway\_station"].fillna(0)

(3) The number of rented houses in the community contains a small number of vacancies, which shall be filled with the average value

mean\_ val = all\_ data["rent\_quantity"].mean()

all\_ data["rent\_quantity"] = all\_ data["rent\_quantity"].fillna(mean\_val)

(4) There are few vacant values of "position" and "area", which are category variables and filled with mode.

mode\_ area = all\_ data["area"].mode()

mode\_ position = all\_ data["position"].mode()

all\_ data["area"] = all\_ data["area"].fillna(int(mode\_area))

all\_ data["position"] = all\_ data["position"].fillna(int(mode\_position))

图形用户界面, 文本, 电子邮件

描述已自动生成

Characteristic engineering

Feature engineering refers to the process of transforming the original data into the training data of the model. Its purpose is to obtain better features of the training data and make the machine learning model approach this upper limit.

Feature engineering can improve the performance of models, sometimes even on simple models.

According to the meaning of the original feature, we build a new feature based on the original feature through feature transformation for later model training and numerical prediction. The specific method is as follows:

a. Total number of bedrooms=number of bedrooms+number of halls+number of bathrooms. In order to avoid linear correlation of four variables, the number of halls is deleted.

b. Proportion of bedrooms: number of bedrooms/number of halls, number of bedrooms/number of bathrooms, number of halls/number of bathrooms

c. House orientation: because the house area is not a number and contains multiple directions, the house area is converted into eight dummy variables, representing whether the house has eight orientations, namely east, west, south, north, northeast, southeast, northwest and southwest

d. Average rent: the houses are grouped according to the name, area and location of the community, and the average group rent is calculated. Three new average rent columns are generated corresponding to the name, area and location of the community

Because some variables are category variables, they are converted to string types. Then, the total data set after preprocessing is divided into training set and test set for model training and prediction.

5. Model Selection

We need to forecast the monthly housing rent based on 17 characteristic values. This is a regression problem. In this paper, we have tried the following different models for training and prediction, and selected several models with better effects for model fusion to get the final model.

（1）LightGBM

（2）XGBoost

The algorithm measures the regression model by calculating the root mean square error of the predicted value and the real monthly rent. The smaller the root mean square error is, the better the regression model is. The calculation formula of root mean square error is as follows:

Where, n is the number of samples; - Forecast monthly rent; X - actual monthly rent;

We bring the training set data into different models to get the prediction results and compare them with the original target data (rent) to calculate RMSE:

6. Result analysis

In this paper, the factors and indicators that affect housing prices are taken as input variables, and LightNGB and XGBoost models are used for modeling. The prediction results are compared with the actual results, and then the model is optimized by considering the characteristic engineering and model parameters such as adjusting the number of input variables, and modeling again to obtain new prediction results. Finally, compare the results of several models, draw conclusions, and interpret the practical significance of the prediction results. By comparing the error between the prediction results of LightGBM model and XGBoost series model, we can find that the error of the prediction results of LightGBM model is the smallest, while the error of the prediction results of XGBoost model is larger, indicating that LightGBM model is better than XGBoost model in the fitting of house prices.

According to the prediction results and actual results of various models, house prices are still on the rise. We should realize that for a stable developing economy, the steady rise of house prices is a normal phenomenon. As long as the growth rate of house prices is basically the same as the growth of GDP and the improvement of people's income, we can think that the real estate industry is operating within a reasonable range, The housing price level is affordable, but the current housing price in China is obviously much higher than the income level of our people. Ordinary wage earners often have to spend their savings for generations to buy a house, which makes people have to reduce their daily consumption to reduce spending, resulting in a decline in market consumption demand. The economic system relies too much on the real estate industry, which will lead to insufficient consumption as a whole and affect the healthy development of the real economy, More young people choose to leave the first tier cities because they cannot afford the high housing prices, and the high housing prices are also a major factor leading to the rising cost of living in the first tier cities. Under such adverse factors, more and more young people will not choose to stay in the first tier cities for development, which will make it difficult for the first tier cities to continue to develop.