Predicting the GDP of China Cities

Pengyang Jiang

October 9, 2020

1. Introduction

1.1. Background

Gross domestic product (GDP) is a monetary measure of the market value of all the final goods and services produced in a specific time period. GDP (nominal) per capita does not, however, reflect differences in the cost of living and the inflation rates of the countries; therefore, using a basis of GDP per capita at purchasing power parity (PPP) is arguably more useful when comparing living standards between nations, while nominal GDP is more useful comparing national economies on the international market. Gross Domestic Product (GDP) per capita shows a country's GDP divided by its total population. The table below lists countries in the world ranked by GDP at Purchasing Power Parity (PPP) per capita, along with the Nominal GDP per capita. PPP takes into account the relative cost of living, rather than using only exchange rates, therefore providing a more accurate picture of the real differences in income.

1.2. Problem

The GDP of a region can be inflected in many aspects, including the urbanization rate and the average life span of the region. In a word, the higher the urbanization rate is, the higher the average life span is, more likely is that the GDP of the region is higher. So, as we have learn to explore the neighborhoods of a city, I wonder if the vary type of the venue in the neighborhood can be a reflection. To be explicitly, if the rate that cafe takes of the total venues is higher in a same cultural environment, the more likely the GDP tends to be higher, because the need for unnecessaries imply that people are more healthy. In my research, I'm going to verify my hypothesis.

1.3. Interest

If my hypothesis works out, I believe there will be two group of audience will be interested in my research. First, the city management official, if they are to improve the GDP of the city, will set some policy which encourage the opening of some venues that may have positive effect to the city's GDP. Secondly, some entrepreneurs may be interested. After the build and calculate the predicted GDP, the will know if the venue they are t open is existing enough in the city.

2. Data acquisition and cleaning

2.1. Data resources

The data of GDP of each city in China is all could be found online, and also the population of residence of each people is also published on the internet on a annual base. I download those data from here. However, I found some problems. For example, the coordination of some city is exist in one list while the GDP of this city is not exist. Then I found those city are too small and most survey neglect those cities. And the newest and the most precise data I used is 2018's GDP.

2.2. Data cleaning and feature selection

The total data have 2255 samples and 205 features. There are 205different venue type included, therefore, we could clearly claim that there would be a lot of redundancy in the feature. For example, the city Panzhihua City, it has the only juice bar while any other city does not have a juice bar. I believe only a sample won't contribute to the GDP while in the model, if not deleted, this sample will have a unpredicted impact on the model output.

So base on the number of each venue type, 78 kinds of feature are deleted. The following shows some feature that are kept and deleted.(Table 1)

Table 1. Simple feature selection during data cleaning.

Kept Features	Dropped features
Hotel, Coffee Shop,	Afghan Restaurant
Fast Food Restaurant,	Mediterranean Restaurant
Shopping Mall,	Bike Rental / Bike Share
Train Station,	Huaiyang Restaurant
Park,	Taiwanese Restaurant
Chinese Restaurant,	Shanxi Restaurant
Pizza Place,	Boarding House
Bus station,	Indonesian Restaurant
Historic Site	Pastry Shop
``	```````

Besides that, there is another problem. Since Foursquare is not popular in China, in some small cities, only a few venues could be explored. For example, in Nanchong City, only four venues could be found when the distance reaches 20000. To solve this problem, I will filter those whose venues are more than 10 and 130 cities remains as samples.

3. Exploratory Data Analysis

3.1. The GDP of each city

We could look into the Figure 1 and find the conclusion that most cities' GDP per capita concentrate between 30000 and 75000. Some cities GDP per capita is more that 150000 RMB.

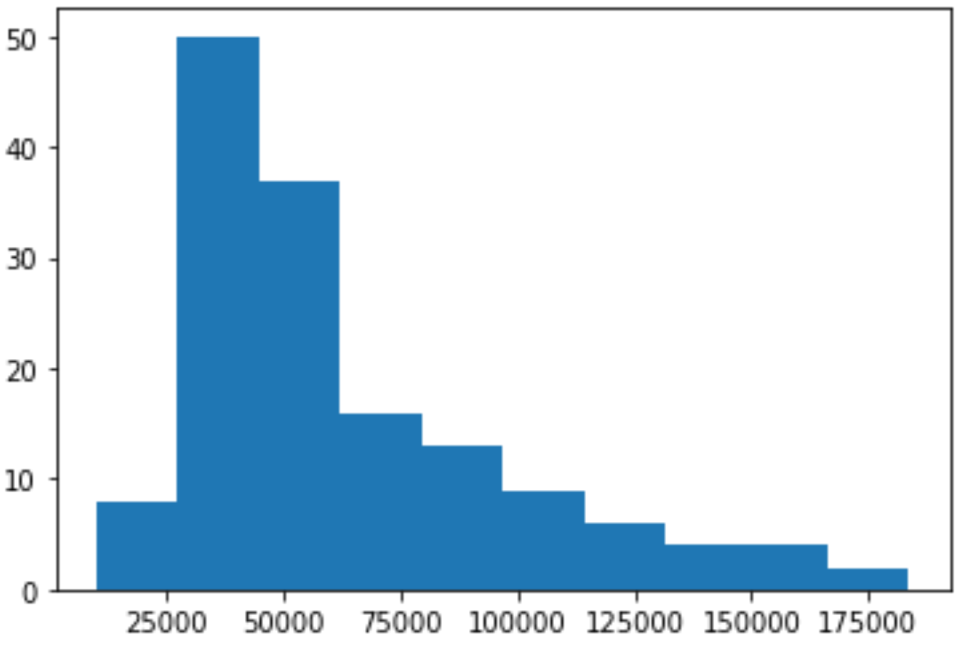


Figure 1. Frequency histogram of GDP per capita of cities in China

3.2. The relationship between feature and cities.

After examining the data, the type feature I used come out to be weak feature, which mean when we look into a single feature, the relationship remains unclear. We should look right into the model building.

4. Predicting GDP

There are two types of models, regression and classification, that can be used to machine learning. For that in this research we are trying to predict the GDP per capita, we should use the model of regression to build the model. In the following, we will use two kinds of model and the second model is used base on the reflection of the first model.

4.1. Linear Regression

I applied linear regression model to predict the GDP per capita of the tested cities. The results had a very serious problems. The predicted values had much narrow range than the actual values (Figure 2), and as a result, the prediction errors were larger as the actual values deviated further from zero (Figure 4) and the score of the model hit the negative. The results were not acceptable because the model does not reflect the predictive ability.

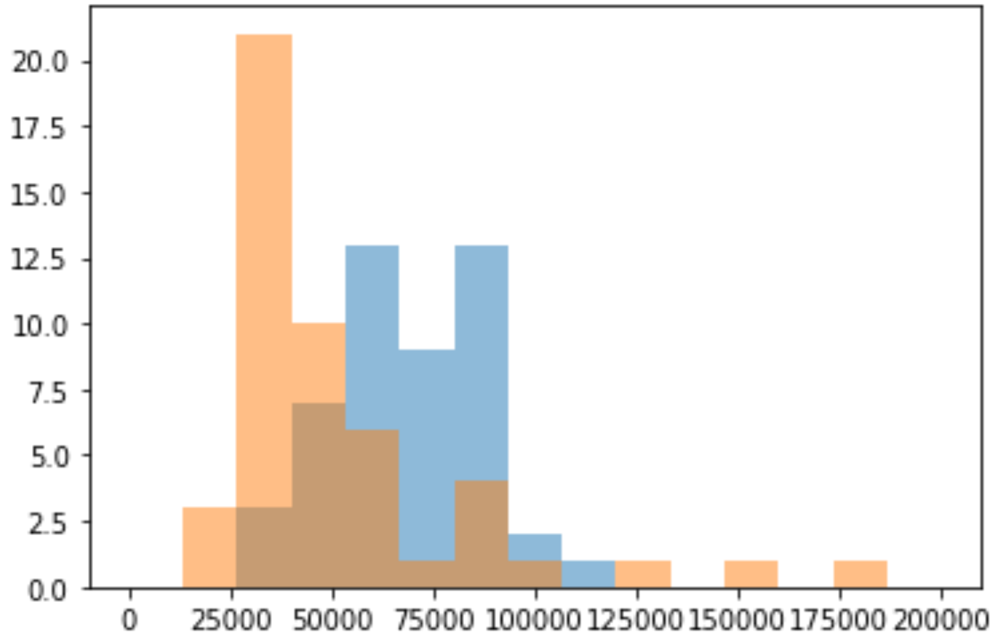


Figure 2. Distribution of actual and predicted GDP per capita using linear regression with equal weights of samples.

The predicted the GDP per capita is deviated from the actual GDP per capita.(Figure 3)

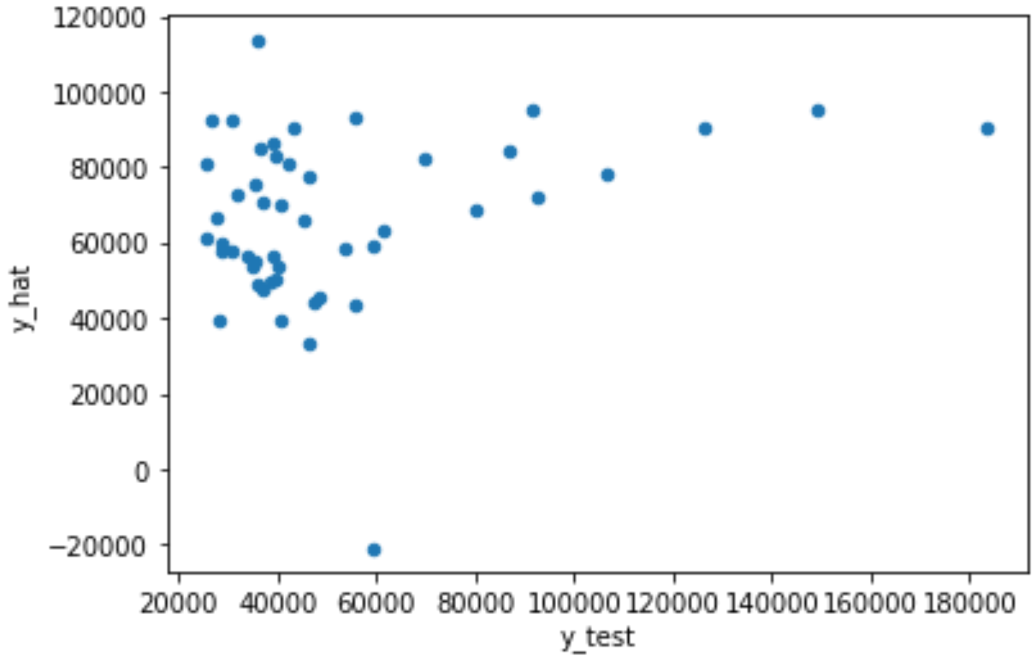
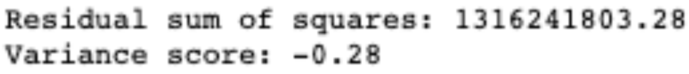


Figure 3. Scatterplot of actual and predicted GDP per capita of test group using linear regression.



```
Residual sum of squares: 1316241803.28
Variance score: -0.28
```

Figure 4. Residual sum of squares and variance score using linear regression.

4.2. GBDT Regression

Considering those features are weak features, the utilize of linear regression model could cause lots of problems like overfitting. Therefore, for better modelling, I used GBDT algorithm to build my model.

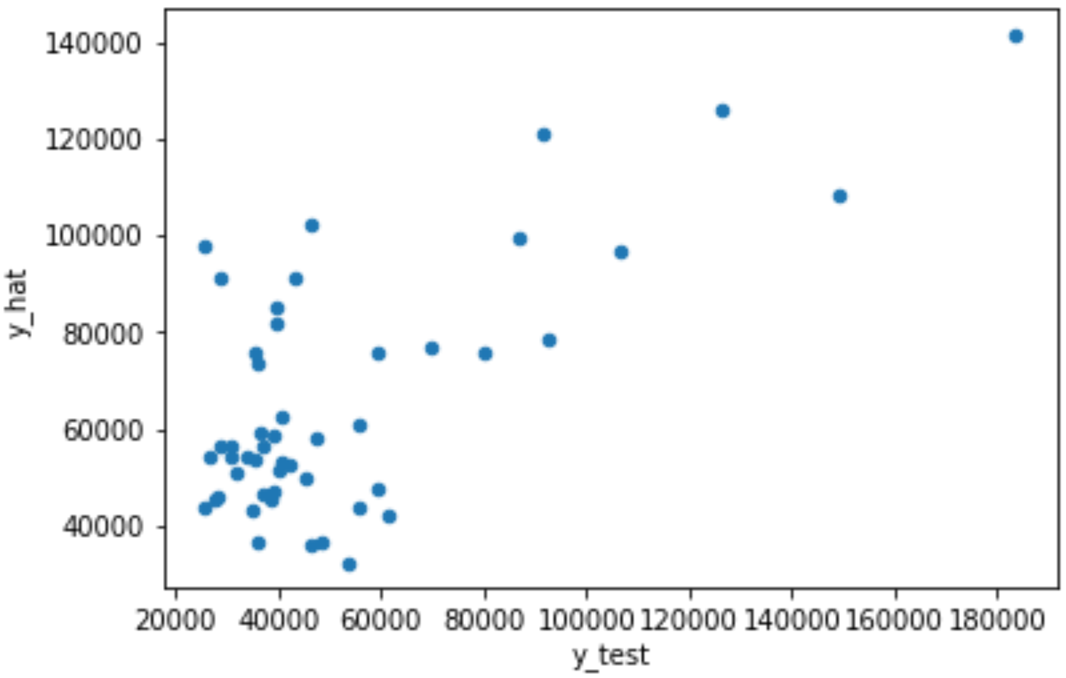
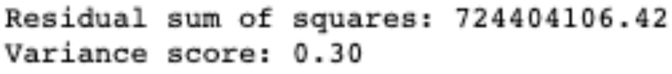


Figure 5. Scatterplot of actual and predicted GDP per capita of test group using GBDT regression.



```
Residual sum of squares: 724404106.42
Variance score: 0.30
```

Figure 6. Residual sum of squares and variance score using GBDT regression.

5. Conclusions

In this study, I analyzed the relationship between GDP per capita of cities in China and the local venues data. I identified age, win share, minutes/games played, improvement last season among the most important features that affect a player’s improvement next season. I built both linear regression model and GBDT regression model to predict the GDP per capita of cities in China. The latter can help to help predict economy performance in a number of ways. For example, the city management official, if they are to improve the GDP of the city, will set some policy which encourages the opening of some venues that may have positive effect to the city's GDP.