Final Report of Machine Learning Project – Apartment Rental Price Prediction

Hao Ge, Zizhuo Liu, Xu Wang

1 INTRODUCTION OF TASK

1.1 The definition of the task

Our task is to help students in Chicago area determine a reasonable price to sublease their apartment or find a sublease via machine learning approach. In this project, the input are attributes of the apartment subleased, along with other factors such as sublease period. A simple illustration is shown in Figure 1.

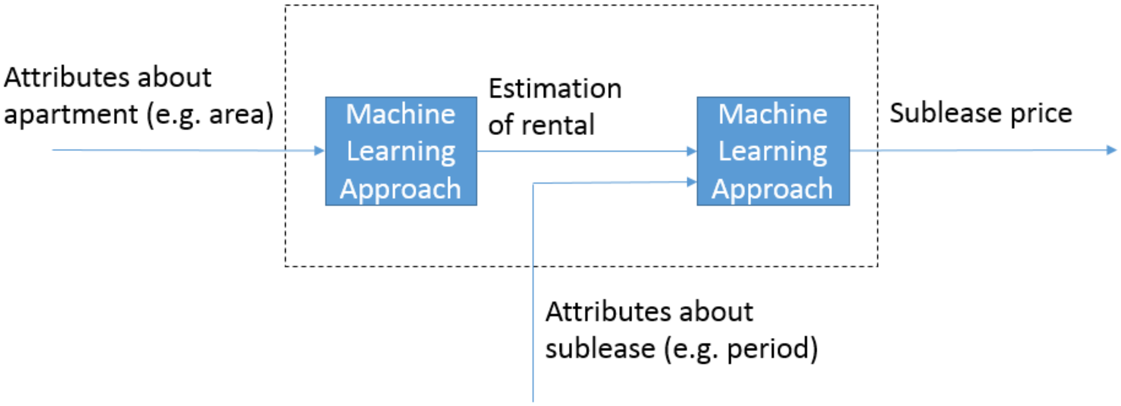


Figure 1: A simple illustration of our task. The inputs are attributes about the apartment and the sublease. The output is the price.

1.2 The meaning of the task

Nowadays especially in the summer, a great number of students might leave for other places temporarily for exchange or internship. In these cases, students often consider to sublease their own apartments and look for a sublease in the city they will move to. Our task is significant for these students, who need to determine the reasonable price and find a suitable sublease.

2 INFORMATION OF DATA-SET

We collect data about rental from Craigslist. A web spider is written to help us find and organize the data on the Craigslist website. Our program can get the house/apartment rental data in a particular area. The attributes we use are listed in Table 1. Since the data on Craigslist may not be complete for each of the posts, we use '?' to denote the missing attributes. We can use the same method in PS2 to deal with the missing attributes. We grab about 4100 set of data from Craigslist for the project in total.

Attribute	Data Type	Attribute Explanation
Price	Float	The listed rental price
Bedroom	Float	The number of bedrooms
Bathroom	Float	The number of bathrooms
Area	Float	The area of the listed house/apartment
House Type	Nominal	0 →Apartment, 1 →Condo, 2 → House
Cat	Nominal	0 → No cats, 1 → Allow cats
Dog	Nominal	0 → No dogs, 1 → Allow dogs
Parking	Nominal	0 → No parking, 1 →Street parking, 2 → Garage
Dishwasher	Nominal	0 → No dishwasher, 1 → Has dishwasher

Table 1: Attributes list

Moreover, we collect data about sublease from WildcatPad (http://www.wildcatpad.com/) and BBS of Northwestern (http://bbs.nwucssa.org/). Based on the advertisements, we can get information about attributes listed in Table 1 except the rental (most people post sublease price only), therefore we sent surveys to students who posted the sublease advertisements and asked them for the original rental. Till now more than half have responded to us. In addition to attributes mentioned above, we consider more attributes such as move-in and move-out date, utility and number of roommates.

3 RESULT

As shown in Figure 1, our task is divided into two parts: first, we need to get an estimation of the rental; second, based on previous step, we need to predict the sublease price.

3.1 Estimation of Rental

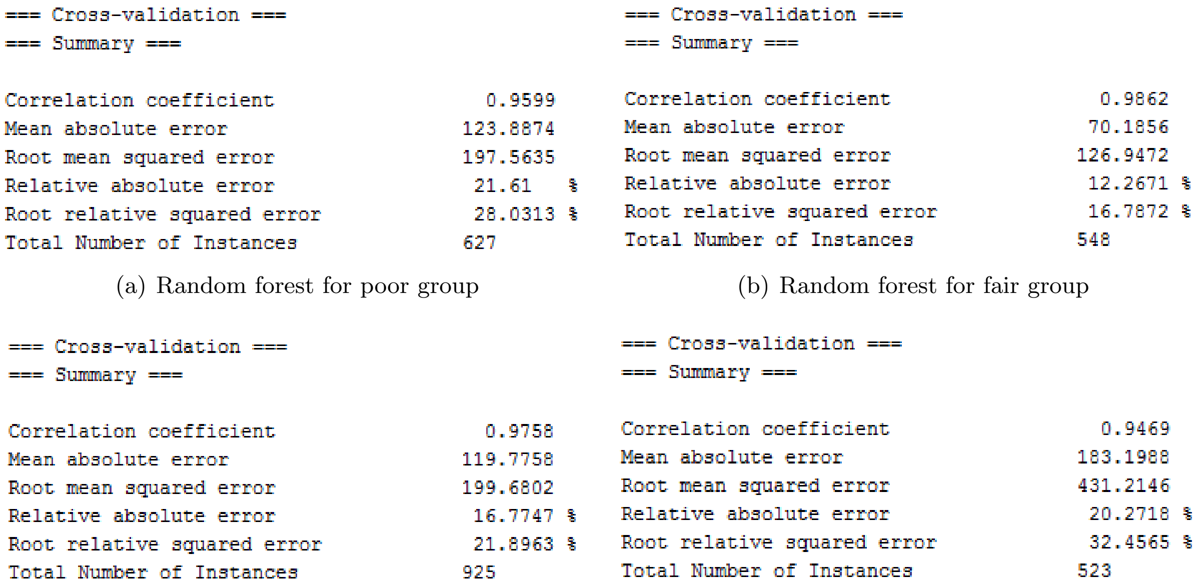
Previously we've worked out the intuitive linear regression model and obtained formula with the form: $\text{price} = 296.6014 * \text{bedroom} + 694.7589 * \text{bathroom} - 0.03 * \text{area} + 384.2259 * \text{cat} + 138.4232 * \text{dog} - 212.2785 * \text{housetype} + 44.3042 * \text{parking} + 373.4674 * \text{dishwasher} + 76.0938$. The root relative squared error was 60.9023%, which was far to be used as a good model.

Since linear regression gives very large error. We use a substitute method for linear regression called regression tree. The principle is to first divide the data into small clusters and then in each small cluster linear regression is applied. Under this method, the relative squared error reduced to around 47%.

And now we've modified our assumptions in different ways to figure out a better model to predict the price.

One solution to increase the accuracy is to build the prediction by dividing the instances into different groups and convert the regression problem to a classification problem so that we could ultilize decision trees to better divide the instances. This is a quite legitimate simplification, since in the terms of rent of the sublease, getting a range of the price should be enough to work as the guide line. By training the model to increase the accuracy, we determined to train the model with the aim of the rent per room (deviding the price by the room number). To cooperate the model with the case of studio instead of the regular apartment with bedroom, we assume the studio has 0.75 bedroom (since the studio is considered as the smaller version of one bedroom apartment). And we round the price per room into the multiples of 200 dollars, which might be an audacius assumption. Because we will also consider the sublease instead of the full price, 200 dollars variation of the predicion would also be smaller especially in the apartment with many rooms. Using all of thse assumptions, we trained our data with Random Tree model and realize the accuracy of 80.33321%. The size of the tree was 595. Considering the complexity of the problem, we think the model worked quite good.

One alternative approach to increase the accuracy is to estimate the hidden attributes based on price per square feet. In details, in addition to attributes listed which could be collected in Craigslist, there are some other important attributes such as decoration, environment around and so on. To compensate for these hidden attributes, we divide the instances into four groups (poor, fair, good and extravagant) based on price per square feet, which is closely related to those hidden attributes. The price per unit square ranges from 0.5 to 7 and roughly follows a Gaussian distribution, which is consistent with the real market. For each group, we implement random forests and linear regression, respectively, as shown in Figure 2 and 3. One interesting observation is for both methods the errors of poor and extravagant group are greater than that of fair and good group, this is because data belonging to the poor and extravagant group lie in the 'tail' of the Gaussian distribution, hence less data is collected and the perplexity is larger.



```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9599
Mean absolute error                    123.8874
Root mean squared error                197.5635
Relative absolute error                 21.61    %
Root relative squared error             28.0313 %
Total Number of Instances              627
```

(a) Random forest for poor group

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9862
Mean absolute error                     70.1856
Root mean squared error                126.9472
Relative absolute error                 12.2671 %
Root relative squared error             16.7872 %
Total Number of Instances              548
```

(b) Random forest for fair group

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9758
Mean absolute error                    119.7758
Root mean squared error                199.6802
Relative absolute error                 16.7747 %
Root relative squared error             21.8963 %
Total Number of Instances              925
```

(c) Random forest for good group

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9469
Mean absolute error                    183.1988
Root mean squared error                431.2146
Relative absolute error                 20.2718 %
Root relative squared error             32.4565 %
Total Number of Instances              523
```

(d) Random forest for extravagant group

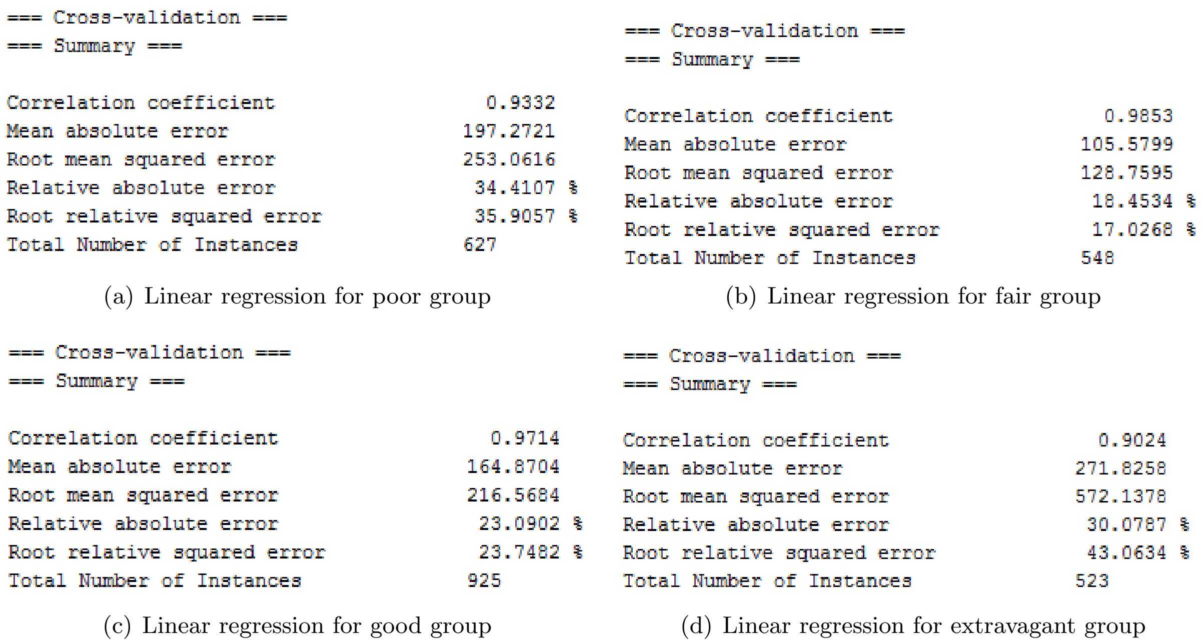
Figure 2: Random forest for four groups

3.2 Estimation of Sublease Rate

We sent surveys to those who posted subleases on WidlcatePad and BBS of Northwestern. To increase accuracy, feedbacks are divided into different groups by the number of bedrooms. Linear regression model is applied in each group and for studios, we assume the number of bedrooms is 1. The relative absolute error is around 24% for different groups. Moreover, we find, compared with the rental, the duration has less impact on the sublease rate.

4 Author Contributions

- Zizhuo Liu worked on the linear regression and random tree model to classify the data sets.
- Xu Wang is responsible for data collection and website design.



```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9332
Mean absolute error                    197.2721
Root mean squared error                253.0616
Relative absolute error                 34.4107 %
Root relative squared error             35.9057 %
Total Number of Instances              627
```

(a) Linear regression for poor group

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9853
Mean absolute error                    105.5799
Root mean squared error                128.7595
Relative absolute error                 18.4534 %
Root relative squared error             17.0268 %
Total Number of Instances              548
```

(b) Linear regression for fair group

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9714
Mean absolute error                    164.8704
Root mean squared error                216.5684
Relative absolute error                 23.0902 %
Root relative squared error             23.7482 %
Total Number of Instances              925
```

(c) Linear regression for good group

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9024
Mean absolute error                    271.8258
Root mean squared error                572.1378
Relative absolute error                 30.0787 %
Root relative squared error             43.0634 %
Total Number of Instances              523
```

(d) Linear regression for extravagant group

Figure 3: Linear regression for four groups

- Hao Ge worked on random forest model and relationship between rental and sublease rate.