CSE 258 Final Project - Two Sigma Connect: Rental Listing Inquiries

Shanyu Chuang (A53104042)
Shiunn An Lu (A53087777)
Hao En Sung (A53204772)

January 12, 2018

Abstract

While searching for a new place to live, it can take hours to look through endless lists on different rent websites. If rent lists can be sorted by each entry's quality, we might be able to find our ideal place to stay in a short time. Our dataset comes from the competition held by Two Sigma and Rentalhop on Kaggle [1]. And our target is to predict people's interest level to each sample on the rent list.

I. Dataset

Our dataset has 49352 samples, and 15 features for each sample.
The features are as following:

- bathrooms: number of bathrooms
- bedrooms: number of bathrooms
- building_id
- created
- description
- display_address
- features: a list of features of the room
- latitude
- listing_id
- longitude
- manager_id
- photos: a list of photo links.
- price: in USD
- street_address
- interest_level: this is the target variable. It has 3 categories: 'high', 'medium', 'low'

First, we plot the distribution of data of each interest levels. We found that 69 percent of the samples have low interest level, 23 percent of the samples have medium interest level, and only 8 percent of the them have high interest level. 1

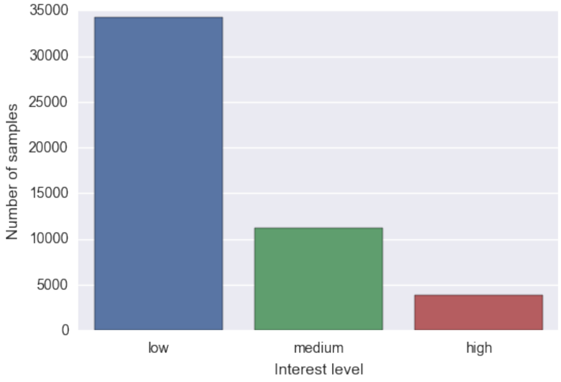

Figure 1: *distribution of interest level*

Second, we plot the distribution of price and found that there are outliers. 2 3

Then we plot the relationship between number of bathroom and the interest level as figure 4. This plot shows that most samples has only 1 bathroom. However, the sample with high interest level has slightly less samples with 2 bathrooms and some of the sample with low interest has 3 bathrooms.

We also plot the relationship between number of bedroom and the interest level as figure 5. We can find the trend that there's a lot of sample with low interest level has only 1 bedroom.

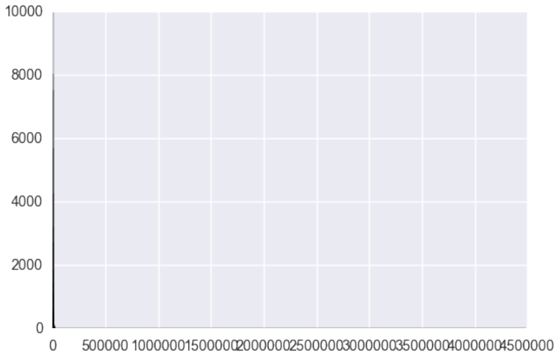

Figure 2: *distribution of price shows outliers*

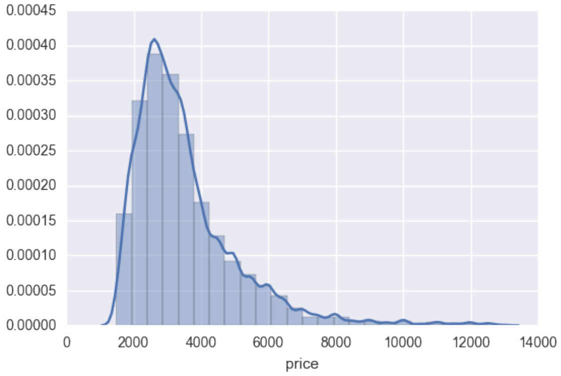

Figure 3: *distribution of price without outliers*

Geographic location is definitely an important feature for our training task. We plot the distribution of longitude and latitude as fig 6 and 7. These two figures show that most of the locations are around the New York City area. Also, we plot the locations of all the samples on base map as 8. On 8, blue samples represent low interest, green samples represent medium interest, and red sample represent high interest. We also found that the some of the samples are in New York City and some are in suburban area. We assume this might be an important factor that affect people's interests.

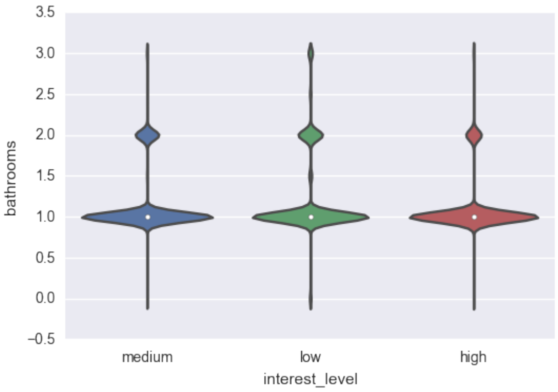

Figure 4: *relationship between number of bathroom and the interest level*

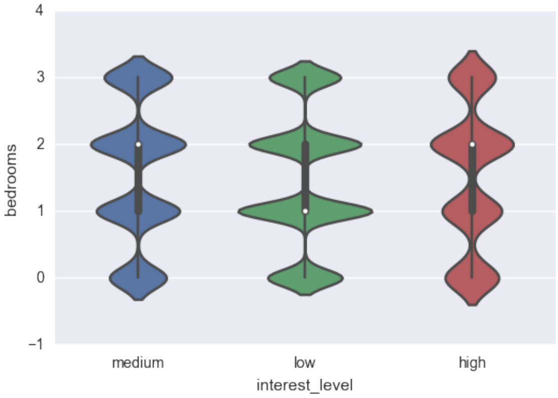

Figure 5: *relationship between number of bedroom and the interest level*

II. Predictive Tasks

i. Evaluation

Our task is to predict the interest level of each sample in test set. The evaluate metric we used is logloss.

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} log(p_{ij}) \qquad (1)$$

Our baseline model is predicting the probability of interest levels by the ratio of samples in the training set. This gives a score of 0.79075 on Kaggle's leader board. We will split our train data into training set and validation set and use validation set to validate our model.

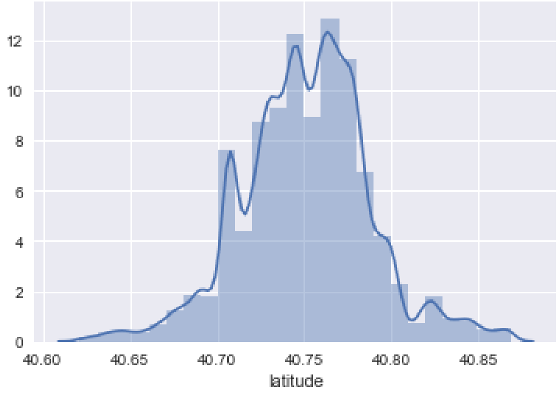

Figure 6: *distribution of latitude*

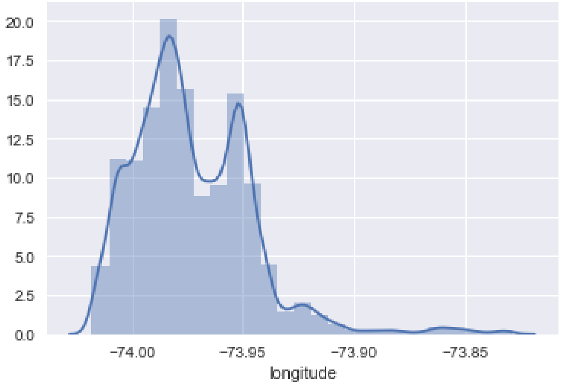

Figure 7: *distribution of longitude*

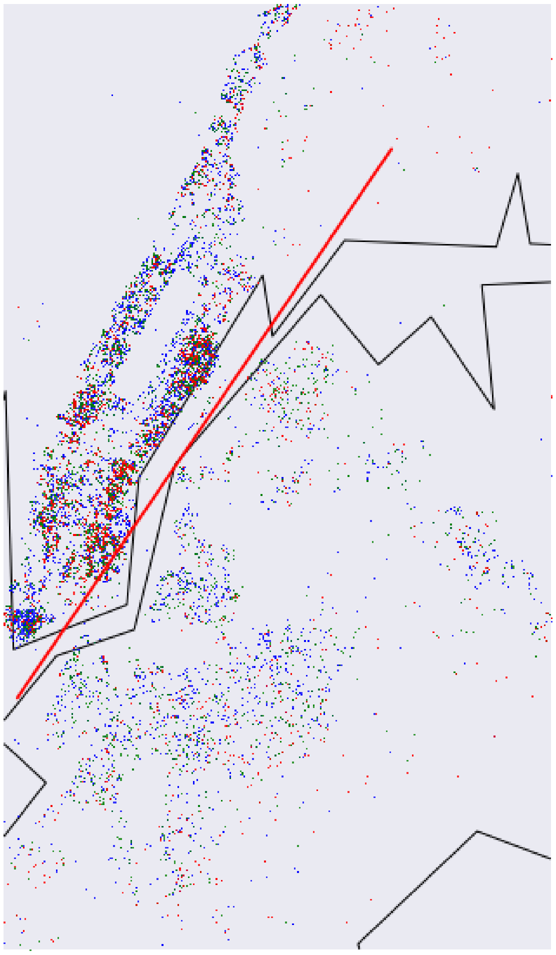

Figure 8: *location of all the samples*

ii. Feature selection

We found that there are duplicate samples in both training set and testing set. We dropped these samples to avoid negatively affecting our model. The features we are using includes: ’bathrooms’, ’bedrooms’, ’latitude’, ’longitude’, ’price’, period of created time in hour, month, and week, ’n_features’, ’n_words_in_description’, ’n_images’, ’building_id’, ’price_diff’, ’price_ratio’, ’city’.

Basic Features include the first 5 features and these features are the original features in training set.

Time Features are extracted from the feature ’created’ which represents the created time of a sample. We use hour 0 to 23, week 1 to 4, as features. Also, we found that all the samples are created within four months so each month is a feature respectively.

Text Features include ’n_words_in_description’, ’n_features’, and ’n_images’. As for number of words in review, ’n_words_in_description’, we have reviews of the string type. We eliminate punctuations and split the strings to obtain the number of words for each review as n_words_in_review. To obtain number of features, we get the list of feature from training set and use its length as n_features. We do the same for number of images, ’n_images’.

ID Leakage Feature indicates ’building_id’. We found an interesting relation between ’building_id’ and interest level. Most of the samples whose 'building_id' is 0, also has low interest level.

The feature 'price_diff', 'price_ratio' is the difference of price with the median of all prices and the price divided by the median of all prices respectively. For 'city', we find a line which can approximately separate city area and suburban area. With this line, we can determine if a sample is located in city or suburban by using cross product to see which side of the line it is located.

III. Methodology

For this specific task — predicting whether people have low, medium, or high interest in certain housings, we propose two different models for learning: classification and regression. We first used several classification models to predict the probability of each interest level of each sample. Then we used regression models by setting low as label 0, medium and high as label 1. We will discuss each experiment as following.

i. Classification

Classification models are used to differentiate different instances into different categories, more often than not, in terms of probabilities.

i.1 Motivation

Since there are three different levels of interests: 'low', 'medium', and 'high', a naive learning strategy is to regard these interests as target labels and model this task as a multi-class classification problem.

i.2 Models

During the competition, we have tried several different models, which can be roughly categorized into two groups: SVM-based model and tree-based model.

- Linear SVC
- SVC
- Random Forest Classifier
- Gradient Boosting Classifier
- XGBoosting Classifier

The first and second model are based on the same policy to separate the data points. The only difference between them is that the former one uses a linear boundary to separate instances; while the latter one utilizes radial basis function (RBF) kernel, which can theoretically maps features from low dimension space into infinite dimension space. Some of the strengths of SVM based methods are 1.it has a regularization term to avoid overfitting. 2. the model only take vectors closest to the decision boundary into considerations. It is good for separating two classes yet it fails to consider obvious cases which can also have insightful information. Another weakness it has is that the performance relies heavily on the choice of kernels.

The following three models are all tree-based models. Random forest model simply builds up multiple decision trees with different portion of data and merge their decisions as final predictions. Gradient Boosting Model is another well-known tree-based model which has strong fitting power but with low risk of overfitting. XGBoosting model is a new boosting model, which is widely used in Kaggle competition. We refer to many others' codes in Kaggle forum and used it as our main model because of its stable performance. The advantages of using tree-based models are handling large set of features without sacrificing too much of the performance, and giving an estimate of important features. Some limitations of tree-based algorithms are decision trees used in these models cannot learn exactly many linear decision functions in the finite case [2].

ii. Regression

Regression models are typically used to predict continuous real values.

ii.1 Motivation

Though 'low', 'medium', 'high' at first glance are three independent categorical features, they in fact have numerical dependencies, i.e. 'low' < 'medium' < 'high'. It is then intuitive to transform them into numerical values and apply regression models onto this task.

ii.2 Strategy

There are many different ways to transform 'low', 'medium', 'high' into numerical values. For now, we naively regard 'low' as 0, 'medium' as 1, and 'high' as 1. In this way, we can use regression model for this problem.

For the prediction part, if an instance is predicted as value x, we map it back into an array of three probabilities for 'low', 'medium', and 'high', with following steps.

1. Normalize x into range $[0, 1]$.
2. Use $(1 - x)$ as the probability of being 'low'.
3. Use $x \times r$ as the probability of being 'medium'.
4. Use $x \times (1 - r)$ as the probability of being 'high'.

The parameter r is now set as 0.75, which is the ratio between number of 'medium' and number of 'medium' + number of 'high' among instances.

ii.3 Models

Similar to classification models, we experimented the following five regressors and used *XGBoosting Regressor* as our main model.

- Linear SVR
- SVR
- Random Forest Regressor
- Gradient Boosting Regressor
- XGBoosting Regressor

IV. Literature

In this competition, we predict how popular an apartment rental listing is based on the listing content like text description, photos, number of bedrooms, price, etc. The data comes from renthop.com, an apartment listing website. These apartments are located in New York City.

Our dataset is from renthop and it has not been studied in other research. Studies using similar datasets related to housing and apartments have similar features such as number of rooms, square foot, and price.

i. Related Work on Similar Data Sets

In [3], the author combines both visual and textual attributes to be used for price estimation. The collected dataset is composed of 535 sample houses from California State in the United State. Each house is represented by both visual and textual data. The visual data is a set of 4 images for the frontal image of the house, the bedroom, the kitchen and the bathroom as shown in figure 1. The textual data represent the physical attributes of the house such as the number of bedrooms, the number of bathrooms, the area of the house and the zip code for the place where the house is located. They used both SVR and multilayer neural network to estimate prices. Through experiments, they found that aggregating both visual and textual information yields better estimation accuracy compared to textual features alone.

In [4], the author proposed a new type of weighted support vector regression (SVR), motivated by modeling local dependencies in time and space in prediction of house prices. The weight functions in the house pricing model depend on the geographical distance to the house of interest and the difference in time of sale (CF-weights) as well as the differences lying in variables (OF-weights), such as house size and number of floors. The results illustrate that the combination of the two types of weights describes the relative importance of observations very well and lowers the influence of possible outliers.

In [5], the author compares the predictive power of the hedonic model with an artificial neural network model on house price prediction. A sample of 200 houses in Christchurch,

New Zealand is randomly selected from the Harcourt website. Factors they used including house size, house age, house type, number of bedrooms, number of bathrooms, number of garages, amenities around the house and geographical location are considered. Empirical results support the potential of artificial neural network on house price prediction.
Physical characteristics of real property are considered as the deciding factors for housing value. However, values calculated on the basis of only property characteristics may not give accurate predictions. This can be due to ignorance in considering features that indirectly affect the values. In [6], the author identified and presented variables in two categories, namely intrinsic variables and extrinsic variables. This can be used as a reference for us to choose appropriate features in our prediction.
In [7], there are 2 NYC students participated in a Kaggle contest of house price prediction.

ii. Summary

The preferred models in the studies of housing are support vector regression and neural networks. Using both visual features and textual features will get better performance than only using textual features.

V. Result

i. Classification

Among five models we tried, i.e. Linear SVC, SVC, Random Forest Classifier, Gradient Boosting Classifier, and XGBoosting Classifier, XGBoosting Classifier has the best performance with parameters as follows.

- n_estimators=1000
- learning_rate=0.1
- max_depth=3
- min_child_weight=1
- colsample_bytree=.9
- colsample_bylevel=.5
- gamma=0.0005
- scale_pos_weight=1
- base_score=.5
- reg_lambda=1
- reg_alpha=1
- missing=0
- seed=514

After fixing the model, we spent most of the time in feature engineering. Our model would have significant improvement each time when we find out some really informative features.

Our trials with significant improvement on public scoreboard can be listed as follows. It can be found out that XGBoosting classifier with increasing features are robust against overfitting and it can produce better results.

Model	Score
Baseline: global ratio	0.79075
Basic Features	0.76980
Basic + Time Features	0.62827
Basic + Time + Text Features	0.60554
Basic + Time + Text Features + ID Leakage	0.59283

Table 1: *Features Effectiveness with Leaderboard*

ii. Regression

Due to the various ways to transform level of interests into numerical values and different strategies to adjust the ratio r, we do not spent too much time diving into the model fitting. In our best model trial, we can reach around 0.61 in 10-fold cross-validation, which is competitive to our experiment results with classification model.

On top of that, we share our novel ideas and codes through Kaggle forum and lead some interesting discussion. We are convinced that regression is the best way to deal with this task and we might have better performance with more time investigate in it.

VI. Conclusion

During this assignment, we firstly devote most of efforts in researching features and visualize features. Later, we use classification model to gain around 0.2 improvement (from 0.79 to 0.59) in terms of negative loss error. Beside that, we further propose to numerically represent the labels 'low', 'medium', and 'high' and use regression model to obtain approximately 0.61 score in ten-fold cross-validation.

We also learned a lot from the forum discussion, including understanding other's motivations behind created features and reading through their codes. Furthermore, we share our novel idea of using regression model through forum and trigger many interesting discussions.

References

[1] Two Sigma. Two sigma connect: Rental listing inquiries, 2017.

[2] Dietterich T.G. Ensemble methods in machine learning. in: Multiple classifier systems. *International workshop on multiple classifier systems. Springer Berlin Heidelberg*, 2000.

[3] Mohamed N. Moustafa Eman H. Ahmed. House price estimation from visual and textual features. *NCTA, 8th International Conference on Neural Computation Theory and Applications, At Porto, Portugal*, 2016.

[4] Xixuan Han and Line Clemmensen. On weighted support vector regression. *Quality and Reliability Engineering International*, 30(6):891–903, 2014. QRE-13-0415.R1.

[5] Visit Limsombunchai. House price prediction: Hedonic price model vs. artificial neural network. 2004 Conference, June 25-26, 2004, Blenheim, New Zealand 97781, New Zealand Agricultural and Resource Economics Society, 2004.

[6] Sayali S. Sandbhor and N. B. Chaphalkar. State of art report on variables affecting housing value. *Indian Journal of Science and Technology*, 9(17), 2016.

[7] Chenbo Tang Christine Stelmer and Nicolai Tanghøj. A study of airbnb prices in copenhagen, 2016.