Evaluation of Customer Interest-Level of Rental Listings

Xialiang Liu, Yihui Wu | NYU CIMS

Contributions

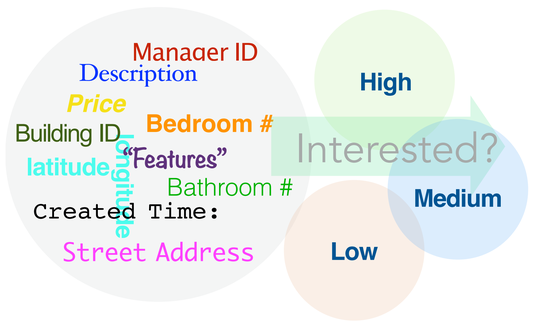
In this project, we build a predictive model of how popular an apartment rental listing is associated with the listing content. Good assessment of such listings will allow owners and agents to better understand renters' needs and preferences, and in verse make it easier for renters to find comfortable apartments.

Dealing with this noisy dataset with expected outliers, we have investigated on tree ensemble methods, *Random Forest* and *Boosted Trees*, to improve our learning model from traditional algorithms of multiclass classification.

Efforts on feature selection and data cleaning for vocabular features play an important role in interpreting the factors appealing to customers.

Problem Settings

This dataset from RentHop (*www.renthop.com*) contains 49352 listings of apartments in New York City. Our objective is to give an estimation based on the available information on the rental listing websites, of the number of inquiries a listing has in the duration that the listing was available on the site, and label it with 3 categories, "high", "medium", and "low". The data involves multiple features ranging from numbers, dates, categorical features, text, geographical data, and the links of photos. We are aiming at predict the interest-level with this dataset and figure out what are the renter's main considerations. With progress of training our learning models, we want to improve the probability of predicting the correct class and the accuracy of prediction.



Features

With access to only five numerical features in the original dataset (number of bedrooms/bathrooms, price, latitude and longitude), we have extracted the following features to improve our learning model.

1. Adding Numerical Features

A natural candidate of adding new numerical features is to describe the amount of information. Together with "created time" of the listings, we can extract 6 features:

- Number of photos (1)
- Number of features (1)
- Number of description words (1)
- Created month, day, and hour (3)

2. Categorizing Hashable Features

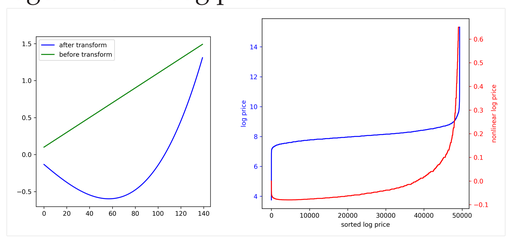
ID numbers that associated with listings, can be encoded into categorical features:

- Address: displayed and street addresses (2)
- Manager ID number (1)
- Building ID number (1)
- Listing ID number* (1)

There will be further discussion on the manager ID and listing ID numbers.

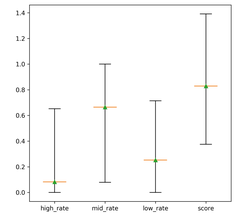
3. The Role of Price

A listing with extremely low or high price is very unlikely to attract the interests from customer. We have incorporate an non-linear transformation on "price" as a complement, which is orthogonal to the log price.



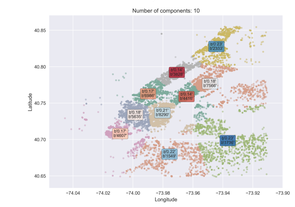
4. Human Factor - The Manager

We look at the performance of mangers, which is defined by the ratios of apartments in each interest-level, under his/her management, and an overall score of the manager is computed based on the three ratios. This set of features is an alternative to categorizing manager ID number. An overview of the manager performance:



5. Clustering on Geographical Data

With the assumption that close locations will have similar community services, convenience of transportation and so on, we will compare the price of each apartment with the median in its "community", which is implemented by clustering on latitude and longitude features. This gap between the price of a apartment with the median price can be regarded as a specific factor of the apartment. (An Example of 10 clusters)



6. What Do Descriptions "Say"?

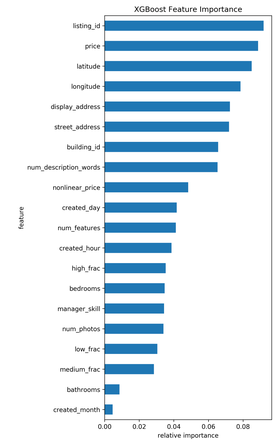
The first step of processing vocabular data is to remove synonyms, as the descriptions in the listings are entered by different managers/officers. Then, we obtain sparse features of tf-idf information through word-map vectorization. Regularization is introduced to avoid overfitting.

Results

After repeated testing, we conclude that the set of features including numerical features (original and added), categorical features, non-linear price, manager scores and the tf-idf features (150-200) gives the overall best performance, with gradient boosting methods. As we adding new features into our model, boosted trees implemented with XGBoost provides with minimal error rate and highest probability of predicting the correct class.

The enormous tf-idf features are not plotted in the figure of feature importance. Top vocabularies includes:

- public_outdoor
- laundry_in_building
- recreation_facilities
- eat
- 24 (24/7 services)
- no_fee
- air_conditioning
- private_parking
- walls_of_windows
- pool



Future Steps

There are room photos on the website for most of the listings, which provides additional high dimensional data. Neural Networks will be a good candidate to learn from the image features.

Information about the convenience of transportation is missing in this dataset, though partly revealed in the descriptions. A combination with urban transportation system and medical systems (i.e. distance to the nearest train station or clinic) might be good to help.

References

[1] Niculescu-Mizil, A., Caruana, R. Obtaining Calibrated Probabilities from Boosting In *UAI '05*

[2] Niculescu-Mizil, A., Caruana, R. Predicting good probabilities with supervised learning In *ICML '05*

[3] Will McGinnis. Beyond One-Hot: Aa Exploration of Categorical Variables. In *Will's Noise*