CSE258 Assignment 2

Bolun Zhang, Moyuan Huang, Yixin Yang, Zhipeng Yan
University of California, San Diego
La Jolla, CA 92093
{boz028,moh,zhy136,yiy235}@eng.ucsd.edu

Abstract

In this report, we

1 Introduction

1.1 Dataset

1.2 Task

2 Model Description

2.1 feature selection

We assessed potential features from the dataset, including categorical features like *feature* and *manger id*, and *todo*.

2.1.1 manager_id

Every apartment posted in Renthop has a related manager which is presented as 'manager_id' in each json entry. We think that this may be ultilized as a bias term for the interest level(similar to the user_id and item_id in assignment1), since different housing manager has different skills in selling apartment, and some of the managers may be usually very attractive to their clients.

We first counted the number of apartments sold by each housing manager, and illustrated the distribution in fig. 1. It is clear that the distribution can provide us with some information about how the managers are doing.

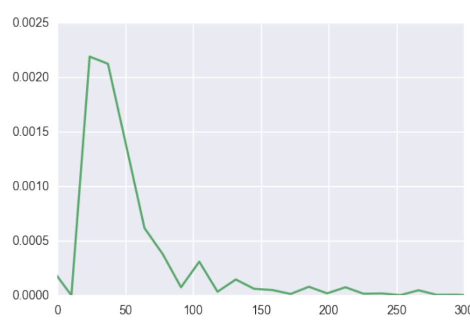

Figure 1: distribution of total number of sold houses

To have more insight about the manager_id, we further computed the fraction of each interest level for each manager, e.g. how many hoses sold by a housing manager is in high interest level? We visualized the the result in fig. 2.

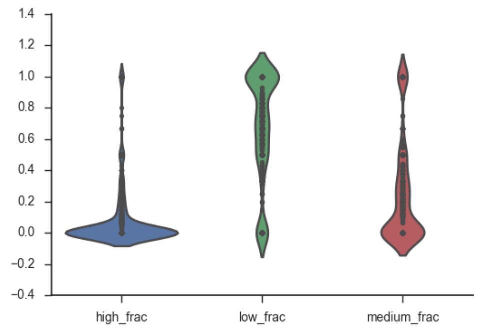

Figure 2: Distribution of different interest fraction.

This statistic have such discriminative ability because an apartment's interest level can be determined by its corresponding manager. The apartment might receive high interest level if the manager has been selling high interest level houses, i.e. the value high_frac of the manager is high.

We thus use this three statistics to represent the quality of the managers, and define *manager_skill* as

$$manager\_skill = 0 * low\_frac + 1 * medium\_frac + 2 * high\_frac$$

After computing the manager_skill, we make a quick evaluation of this metric, which is illustrated in Tab. 1. We can see that the higher the interest level, the more manager skill it requires, which reveiled that our metric actually works. Detailed experiment results can be found in Section 4.

interest level	manager skill
low	0.341852
medium	0.510785
high	0.611761

Table 1: average manager skill of different interest levels.

2.1.2 feature

The feature stores a series of featuring of an apartment, for example, no fees, cats allowed, hardwood floor, etc. All of these could have impact on the interest level of the house. For example, in fig. 3, the interest level has different distribution depending on whehter *dishwasher* is present or not. We then discuss how to include those features in our model.

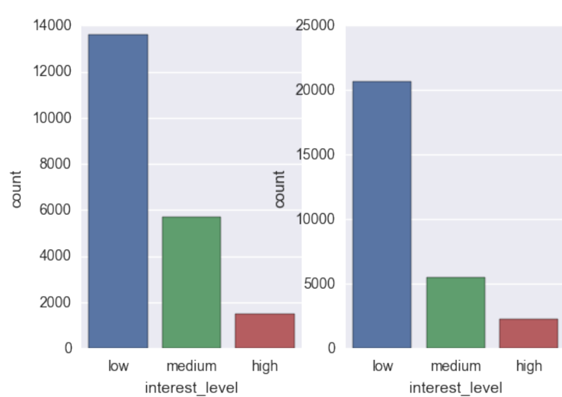

Figure 3: different distribution of interest level with/without dishwasher.

To achieve this, we first discretized those categorical features into separate features and count their appearance across the dataset. The top counted features are listed in Tab. 2. We then used one hot representation of each top counted feature and concatenated them into the original feature set.

feature	count
elevator	26273
hardwoodfloors	23558
cats allowed	23540
dogs allowed	22035
doorman	20967
dish washer	20806
laundry in builing	18944
no fee	18079
fitness center	13257

Table 2: Top counted words.

2.1.3 Other attempts

Besides the features discussed above, we also selected some other features that we think may be informative, which is listed in Tab. 3

name	explanation
num_photos	number of posted photos for the apartment.
num_features	number of features in the apartment feature.
created_year	year posted
created_month	month posted
created_day	dateposted
count	total houses posted by this apartment's manager.

Table 3: additional features

3 Related Work

4 Result and Discussion

model	loss
baseline	1
trivial statistics	0.6514
one hot Apt. features	0.6413
manager skill & count	0.6183

Table 4: Multi-class loss of different models.

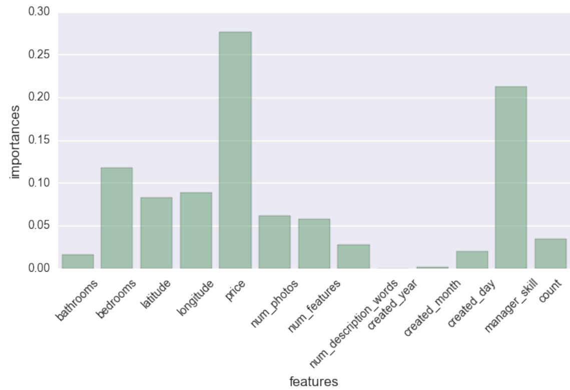

Figure 4: adding *manager skill* and *count*.

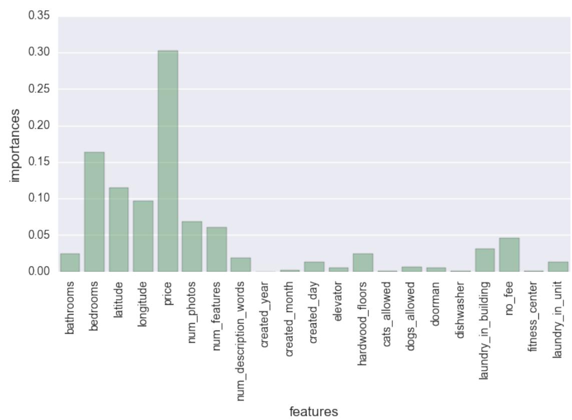

Figure 5: adding *manager skill*.