Evaluation of Customer Interest-Level of House Rental Listings

DS-GA 1003 Course Project Proposal

Machine Learning and Computational Statistics

Xialiang Liu, Yihui Wu

Courant Institute of Mathematical Sciences

New York University

1 Introduction

The first thing after moving into a new city is to find the perfect place to call your new home. However, it is exhausting to browse through endless listings, worry about the accuracy of the description, get confused if some information is missing. There is a great demand of high-quality online posts. However, does those rental agencies really know what is an interesting listing from the renters' viewpoint?

In this project, we will build a predictive model of how popular an apartment rental listing is based on the listing content, in the hope of helping listing website better handle fraud control, identify potential listing quality issues. 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. 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 live on the site, and label it with different categories.

2 Settings

Data

In this dataset, all the apartments are located in New York City. The features involves multimedia data ranging from numbers, dates, categorical features, text, geographical data, and photos. The fields of a rental listing contains are listed in the following table.

The data used in this project is available on *Kaggle* as part of an ongoing competition. The data comes from *renthop.com*, an apartment listing website.

Feature	Description
bathrooms	number of bathrooms
bedrooms	number of bathrooms
building_id	the building id number with alphanumeric characters
created	time of the listing being available on the website
description	short sentences (can be empty)
display_address	name of the avenue/street
features	a list of features/keywords about this apartment (can be empty)
latitude	float number representing its latitude
listing_id	7-digit number
longitude	float number representing its longitude
manager_id	the manager id number with alphanumeric characters
photos	photos of apartments (provided with a link and the corresponding photos)
price	the price in USD
street_address	detailed street information

Outcome

Interest_level: The level of interest is defined by the number of inquiries a listing has in the duration that the listing was live on the site, and has 3 categories: "high", "medium" and "low".

Objective Function

As we will implement different algorithms to search for better solution, we need a assessment criteria for comparison. The performance of an learning model will be evaluated using the *multi-class logarithmic loss*. Each listing has one true class. For each listing, we will obtain a set of predicted probabilities (one for every listing). The loss is computed as,

$$\log loss = -1 \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \log(p_{ij}),$$

where N is the number of listings in the test set, M is the number of class labels (3 classes), log is the natural logarithm, y_{ij} is 1 if observation i belongs to class j and 0 otherwise, and p_{ij} is the predicted probability that observation i belongs to class j.

The predicted probabilities for a given listing resulted from a algorithm might not sum to one, so we will rescale the probabilities. In order to avoid the extremes of the log function, predicted probabilities are replaced with $max\left(min(p, 1 - 10^{-15}), 10^{-15}\right)$.

3 Research Roadmap

Breaking down the three levels of interest "high", "medium", and "low", we can think of the three levels representing whether the building is under-valued, properly-valued, or over-valued by the owner, with the information provided. Thus, to achieve high accuracy, we need to model the expected value of a certain listing. This may include factors on room, location, service, appliances, attractiveness of description, good photos, market price of alternatives, etc.

Progress with Classification Algorithms

Currently we don't know what is the best approach, and we will investigate on different classification algorithms. From the classic algorithms, multi-class **SVM** and/or **logistic regression** will be chosen as the "baseline" algorithms.

However, there are disadvantages with the above algorithms, which leaves the space of improvement. For a two-class problem, if the data is expected to be reasonably clean and outlier free, structural risk minimization is a powerful approach and SVM will work fine. While in a multi-class case and especially outliers are expected, a candidate of alternative algorithms is the **Random Forest**. The subset of training sets with bagging and subsets of features can help reduce their effect.

Besides, inspired by the previous competitions on *Kaggle*, a considerable number of solutions with good performances utilize **boosted trees**. Gradient boosted model (GBM) is a very different type of machine learning models compared with general linear models (GLM), and favors different loss functions. The Gradient Boosted Trees measure the features on different scale, and automatically detects (non-linear) feature interactions. We will experiment with tools such as XGBoost to improve our learning model.

Cooperation of Different Type of Features

This dataset contains multi-media features, which introduces a interesting research on how to combine the predicted results obtained from different types of features. But first we will look at each type of features.

As for the geographical statistics, it can be treated as normal numerical data, but will give up its real-world meaning. Since we don't have an existing feature to indicate convenience of life around the building, we have to create either a score or category of it. Possibly we can preprocess geographical data by **clustering**. Assumption has been made that close locations will have similar community services.

It is likely that we will have three (or four, if the photos can be utilized) from numerical features, language description, and clustered geographical data (and photos if applicable) respectively. The learning model could be benefited from the combination of the above results.

Challenge*: Understanding the Photos

Prediction with images has to training on high-dimensional dataset. We attempt to use feed-forward **neural networks** to process the apartment photos if time permits.

4 Timeline

Mar 08 - Mar 23	Get familiar with the dataset, implement preprocessing and statistical analysis of features. Working on the numerical features and natural language processing first.
Mar 24 - Mar 31	Implement multi-class SVM and logistic regression.
Apr 01 - Apr 12	Experiment with boosted trees and random forest methods, compare with previous results. Look at the cases where the predictive model goes wrong, and research on improvement.
Apr 13 - Apr 19	Improve the predictive model with outputs from different type of features. Reflection and discuss with advisor during the second meet on Apr 19th.
Apr 20 - Apr 30	Continue seeking for improvement of the current model; try to deal with the photos of listings through neural net, e.g. CNN.
May 01 - May 08	Reflect on previous work, understand the principles of each algorithm and compare their favorable situations.
May 09 - May 12	Poster session and final report.