Two Sigma Rental Interest Prediction

January 2020
Columbia University

1. Introduction

The New York based real estate brokerage Fine Housing Inc. makes about 30% of its revenue on the rental market. In a standard procedure for a rental listing, the owner of the apartment asks the real estate broker to find suitable tenants that are willing to pay the desired monthly rent. In return for finding tenants, the owner is willing to compensate the agent with a commission that is customarily equivalent to one or two monthly rent payments.

The main costs an agent and their brokerage firm encounter during a rental process are marketing costs as well as renting below market price and even the risk of losing a rental listing due to failure of finding suitable tenants in the required period of time. Thus, an agent has the incentive to rent out an apartment for a high price due to the proportional commission and also an incentive to rent fast in order to reduce ongoing business costs and minimize opportunity costs. The data set offers 13 columns to build a predictive model for the interest level in a rental listing. The interest level is categorized into three levels: low, medium and high, and corresponds to the number of clicks a listing received in the duration that the listing was live on the site. Hence note that the term "interest" refers to the level of enthusiasm for, as in preference for, a rental, and is not related to the financial concept of interest rate.

2. Data

Data Acquisition

We joined information from three datasets for our project. Our base dataset can be found on Kaggle [1] and includes 13 defining features of rental listings from the website renthop.com including address, price, number of rooms and others. In order to incorporate information about the local crime rate and distance to the nearest subway station, we convert the address first into latitudes and longitudes and then into US zip codes using the "uszipcode" package. The information about transportation convenience was extracted from the second dataset "NYC Transit Subway Entrance And Exit Data" (Auxiliary Data Source I) [2], that includes station name, subway route, latitude and longitude for 1,868 subway entrances and exits in NYC. The relevant crime data was obtained from the third dataset "Crime Complaint Data" from NYPD(Auxiliary Data Source II) [3] whose features include crime complaint number, crime complaint date, latitude and longitude, for 114,673 crime complaints in NYC.

After merging the datasets, we transform the "features" column, that includes especially notable characteristics of each listing into a number of categorical variables. This step required a string separation and comparison algorithm as the datatype of this column was a list of strings.

Data Processing

a. To create the new feature 'distance to subway', we make use of the longitude and latitude of each rental property and connect it to the coordinates of nearby subway stations via Haversine distance [5] (We found this approach to be more efficient than using the Google GeoAPI):

$$d = 2r\ arcsin\sqrt{sin^2\left(\frac{\phi_2-\phi_1}{2}\right) + cos\quad \phi_1\, cos\, \phi_2\, sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)},$$

(where (ϕ_1, λ_1),(ϕ_2, λ_2) is the latitude-longitude coordinate of a rental listing and its nearby subway station entrance/exit, d - distance in km, r ($r_{Earth} = 6371$km) radius of the Earth)

Simultaneously, we obtain the two new features “Route” and “number of subway lines” as byproducts of this step.

b. Utilizing the latitude and longitude data of crime complaints derived from Auxiliary Data Source II, we calculate the total number of crime complaints in each zip code area and merge it with the other features on zip code.

Feature Cleaning

The reason why we need to further clean our features is that after transforming the “features” column in our original dataset into categorical variables, there appear to be more than 1500+ features. Some of them only contain a few observations, which may have no impact on our prediction purpose. Some of them are highly correlated with each other or even the same. Therefore, we need to combine and delete part of features to keep an appropriate amount of features for prediction purpose.

3. Model

Ordinal Regression

The output variable of the model, i.e. interest level, can take on 3 values: low, medium, and high. Hence, it is a multinomial choice model. However, note that the choices are in hierarchical order. Furthermore, note that there is no scale of how much higher “medium” is than “low”, and how much higher “high” is compared to “medium.” This is called an ordinal variable. Other examples of ordinal variables are: rating movies from 1 to 5 stars and describing one’s preference from strongly disagree to strongly agree.
To predict the probability an ordinal variable will take on each of its 3 possible values, we use an ordinal regression model. The ordinal regression algorithm to predict an ordinal variable that can take on k possible values involves doing k-1 regular logistic classifications. Our variable can take on 3 values. So we do two logistic regressions as follows.
The first logistic regression predicts the probability the interest level is above low, i.e. predicts the probability the interest level is medium or high. Accordingly we create a column of data, “HighMedium” with values equal to 1 for all rentals that have high or medium interest level and equal to 0 for rentals with low interest level, and then predict the value of this column.
Similarly, the second logistic regression predicts the probability the interest level is above medium, i.e. predicts the probability the interest level is high. Accordingly we create a column of data, “High” with values equal to 1 for all rentals that have high interest level and equal to 0 for rentals with low or medium interest level, and then predict the value of this column.
Lastly we combine the probabilities Pr(Interest level>low) and Pr(Interest level>medium) obtained in the above two regressions to deduce the probabilities of low, medium, and high interest levels as follows:
1. Pr(Interest level=low) = 1 - Pr(Interest level> low)
2. Pr(Interest level=medium) = Pr(Interest level=[medium or high]) - Pr(Interest level=high)
= Pr(Interest level>low) - Pr(Interest level>medium)
3. Pr(Interest level=high) = Pr(Interest level>medium)

We divide potential factors into two classes: (1) Inherent factors: features that can directly get from Zillow; (2) External factor: features never appear on Zillow.

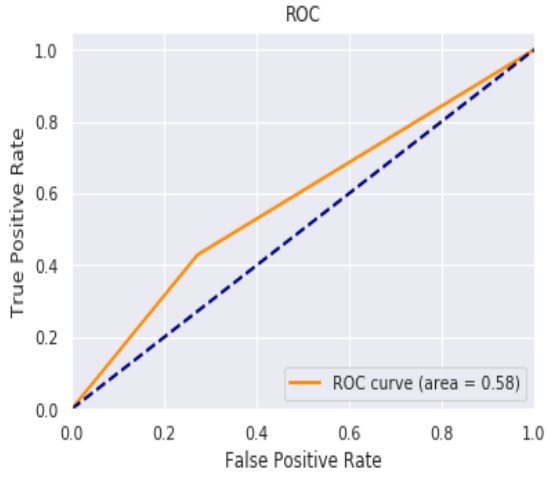
4. Model Result

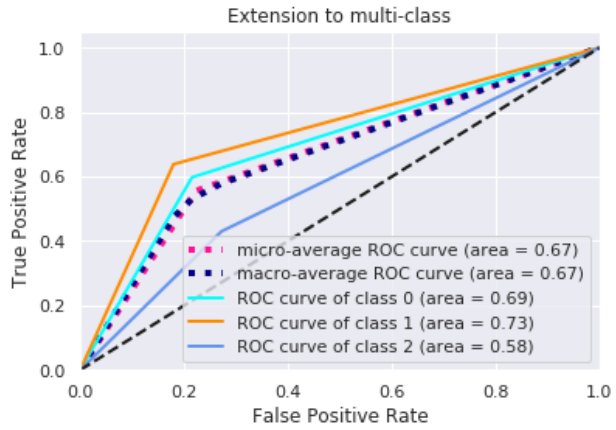
After training our rental interest level prediction model, on the one hand, we introduce true positive(tp), false positive(fp), false negative(fn), true negative(tn), ROC and AUC to evaluate our model prediction.These lead to the following formulas and results:

$$accuracy = \frac{tp + tn}{tp + fp + fn + tn} \quad recall = \frac{tp}{tp + fp} \quad precision = \frac{tp}{tp + fp}$$

$$Negative\ Predictive\ Rate = \frac{tn}{fn + tn} \quad Specificity = 1 - \frac{fp}{fp + tn}$$

$$Positive\ Predictive\ Rate = \frac{tp}{tp + fp} \quad Sensitivity = \frac{tp}{tp + fn}$$





On the other hand, based on the regression coefficient and variable significance statistics, we find the Top 15 influential features that have significant impact on rental interest level (shown in PPT).

5. Business Recommendations

In order to rent out an apartment for a client (customarily the landlord), a real estate agent needs to do a certain amount of marketing to attract as many potential clients as possible. In general advertisement entails costs either in the form of money (direct costs), time or

opportunity costs (indirect costs). The most significant direct sources of costs are publication fees for a rental listing on online platforms and in print media such as booklets or magazines.

Direct Costs:

The company Fine Housing Inc. focuses their online marketing mainly on three websites (assumption, namely New York's most widely used online platforms Street Easy, "Zillow" and their personal website for which they do not have to pay any publication fees.
In addition to these variable costs there are one time fixed costs for publishing ads in print media.

Indirect Costs:

These costs are caused by the circumstance that while an agent is busy looking for clients for a certain listing they will not be able to focus on other endeavors and thus might lose different attractive rental offers. Indirect costs are harder to quantify but will be incorporated in our profit model. The indirect costs also incorporate the negative consequences of a listing staying on the market for an extended period of time. These consequences usually entail maintenance costs and reduced attractiveness of the listing.

Assumptions:

In this section we will list the assumptions that were used to build our business case. It should be noted that these assumptions are not based on data but on qualitative research and common sense. One should keep in mind that these assumptions are the starting point of a sophisticated algorithm and in order to obtain business relevant results the objective should be to verify and adjust these assumptions on the basis of real data analysis.

1) *The average costs to post a listing online amounts to $10 per day.*
   This assumption is based on the fact that publishing a rental listing on Street Easy will cost $6.50 starting from January 2020. Posting the listings on other websites accounts for the remaining $3.50.
2) *The variable cost increases slowly according to a logarithmic increase function.*
   The aim here was to incorporate opportunity costs as well as maintenance costs that occur when a rental listing is vacant for an extended amount of time. To determine the parameters of the logarithmic increase we used common sense deciding that after 10 days the variable costs per day should increase 10% with a decreasing second derivative. The cost per day can be computed by:
   $$C_v(t) = \frac{c_o}{10 * log(10)} * log(t+1) + c_o \ , c_o = \$10$$
3) *There are $150 of fixed costs per publication resulting from the fee to post in print media.*
   $$C_f = \$150$$
   This leaves us with the following cost function:
   $$C(t) = C_f + C_v(t) = \$150 + \sum_{i=1}^{i=t} \frac{\$11}{10 * log(11)} * log(i+1) + \$11$$
4) *The revenue generated by each listing equals one month commission.*
   This is custom on the rental listing market.
   $$R(price) = price$$
   Leaving a Profit formula as:
   $$P(price) = R(price) - C(t)$$
5) *The time that a listing is on the market is inversely proportional to the interest level.*
   We again chose a logarithmic decrease for this function to generate a right-skewed

distribution. Especially, this function would need tuning with the help of data that explains the time a listing was exposed to the market. Maximum Likelihood Estimation of this function via a Weibull distribution could drastically improve the results of our model.

$$t(s) = \left(\frac{180}{log(0.05)} * log(s) + 1\right)$$

6) *The rental interest level is inversely dependent on the price.*
This is rather an observation than an assumption and is the result of our ordinal logistic regression. We create an interest level score out of the results of our ordinal regression and observe the change of interest score when changing the price of a listing.

$$s(p_{low}, p_{medium}, p_{high}) = (1 - p_{low}) * (p_{medium} + 10 * p_{high})/10$$

$$p_{low} \propto price$$

$$p_{medium}, p_{high} \propto \frac{1}{price}$$

We as external contractors offer the following solutions to increase our customer's profit. With our prediction model, we are able to predict the probabilities that a new listing belongs to the low, medium or high interest category. These three numbers are the key ingredients to our business suggestions. We weigh each of these probabilities to obtain an interest score s, which we then translate into cost through various transformations that incorporate the assumptions listed above. As our ordinal logistic regression model uses the price as a regressor to generate the probabilities, we can then analyze how a change in price changes the interest level probabilities and thus the total profit for a specific listing.

In our analysis, we focus on finding the discount factor that represents the percentage change to the proposed price by the landlord. Generally, the customer (landlord) will present a listing offer to a real estate agent with a price that he hopes to obtain from a tenant. Depending on the flexibility of the landlord the real estate agent can suggest minor adjustments to the price in order to reduce the time a listing will remain on the market. This has mutual benefits for the landlord as well as the real estate agent as both of them accumulate costs the longer a listing is on the market.

We use the trained model to find the price suggestion that maximizes the real estate agent's profit while incorporating that a change in price can influence the probability that a landlord offers the listing through "our" vs. the competition channels.

Our model reveals another marketable information that could be used to increase profit. So far most real estate companies that focus on the rental listing do not filter out which listings to take into their portfolio. With the help of our profit prediction we are in a position to evaluate each listing individually and determine whether the object is worth accepting and if not, negotiate a price with the landlord that creates benefits to both parties.

Lastly, the logistic regression reveals which factors influence the interest level the most. Thereby, we can analyze the data provided by the client, select the most relevant features and present it in a way to increase interest level without having to invest any monetary resources. In case, some of the relevant features We also propose a time guarantee model with the objective to ensure that rental listing will be off the market within a fixed period of time. The idea behind this is that generally not only real estate companies have an interest in renting out apartments quickly but even more so do the landlords themselves. An empty apartment has a negative cash

flow due to the fact that there are now rent payments and maintenance costs to keep the rental object attractive. Thus the real estate broker can ask for a bonus from the landlord if they can assure that they will be able to get the listing off the market quicker than the competition.

6. References

- https://www.kaggle.com/c/two-sigma-connect-rental-listing-inquiries/data
- https://data.ny.gov/Transportation/NYC-Transit-Subway-Entrance-And-Exit-Data/i9wp-a4ja
- https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243/
- https://www.cs.waikato.ac.nz/~eibe/pubs/ordinal_tech_report.pdf
- https://towardsdatascience.com/simple-trick-to-train-an-ordinal-regression-with-any-classifier-6911183d2a3c