Predicting inquiry from potential renters using property listing information

Takeshi SO
Daito Trust Construction Co.,Ltd
Institute of Future Design
in Housing Market
Minato ward, Tokyo, Japan
st102456@kentaku.co.jp

Yuta Arai
Reitaku University
AI Business Research Center
Kashwa City, Chiba Prefecture,
Japan
yuarai20@reitaku-u.ac.jp

***Abstract*—In this study, we deduced how accurate the number of inquiries from potential tenants for housing available for rent can be predicted based on the attributes of the housing, using multiple statistical methods, and compared the results. The purpose of this study is to show these results as case studies. Confusion matrices were calculated based on the results deduced with three methods – the classical logistic regression, RandomForest, and XGBoost – and prediction accuracies were verified. The results showed that the accuracy of XGBoost was the highest, followed by that of logistic regression. It is sometimes desirable to use logistic regression, which is easy to interpret from the perspective of application to business, because the differences in accuracy among the statistical methods are not significant. It is thus important in business to take into account the accuracy, ease of interpretation, and research structure and select the most appropriate statistical method.**

Keywords—Housing market, Response Analysis, Property Equipment, Machine Learning, Logistic Regression

I. Research background and purpose

Statistical methods used for data analysis are constantly improving, with numerous researchers developing new algorithms, with novel implementations (functions) being added to statistical software and computer languages.

The more functions implemented, the greater the confusion as to which are best for actual usage. Depending on the purpose of the solution task, general divisions of usable functions exist, including discrete variable estimation, discrimination, variable aggregation, and so on. Nevertheless, the actual selection of appropriate functions often depends on the experiences and strengths and weaknesses of those performing the analysis. In the business field, the trend indicates that clients respond positively when new instead of traditional statistical methods are used.

The background reason for this is that, for a single problem to be solved, the results from multiple functions are being compared on an ad hoc basis, with insufficient practical case studies available for reference. Most of the reports from applied research simply present the data for the problem at hand, the methods (functions) applied, and the results. Very few reports have described why they used certain methods, and how their results would have differed if they had used other methods.

Thus, the current study uses three methods (logistic regression, XGBoost[8], and RandomForest[9]) to derive estimates for the following question: "Can property listing information be used to determine/predict whether potential renters will make inquiries concerning specific listed properties?" Added to this are considerations regarding differences in accuracy among the three methods, and which applications are easiest to apply in actual business, and so on. The purpose is to provide practical research that can be used as a case study.

Accumulation of such case studies can contribute to the selection of appropriate methods for use in practical analyses regarding analytical structure including verification tasks and status, the skills and experience of technical personnel, the number of persons for analysis, and so on.

II. Previous research

A related study regarding the current research question is reported [1][1].

Other studies using predictive models have been conducted in fields other than real estate, for example, reference [2] study on reservation rates for accommodation facilities (hotels, inns, etc.). Predictive model-related research is conducted by a variety

[1] The process before moving into a rental residence can be divided into two broad types. One is where a potential renter first conducts a search by themselves until they find a property suitable for inquiry. The other method is to visit a real estate agent, where detailed information is provided. The potential renter then inspects the property and decides whether to rent it. In the latter process, elements other than advertised information become important. It is therefore thought that it is not enough to use only property information as the explanatory (predictor) variable in model construction.

978-1-6654-1865-2/21/$31.00 ©2021 IEEE
DOI 10.1109/MIPR51284.2021.00053

of private firms for revenue management, such as when companies seek to maximize sales using unit prices and prediction rates. There are many large corporations in the accommodation sector, and these firms have access to the large amounts of data required for analysis. However, conditions and status in the real estate sector differ greatly by country. In Japan, there are almost no real estate companies that have the large amount of data required for robust analyses. This is one major reason for the relative lack of related research.

Much analysis in the real estate sector has focused on estimating rent and prices for condominiums for resale[3-6]. However, classical multiple regression analysis was used in all of these studies, with no comparison made among multiple statistical methods[2].

Reference [7] compared multiple methods, comparing error rates of traditional regression analysis, a neural network, and a regression tree. This study was rich in suggestions, including considerations regarding nonlinear and linear models, and so on.

Nonetheless, there is insufficient research performed with added considerations from the practical perspective of which methods are most easily applied in actual business practice.

III. Research Method

3.1 Data

The present study analyzed rental properties for sale (i.e., listed) from real estate agencies specializing in rentals[3] in the Tokyo Metropolitan area and in the three prefectures of Saitama, Chiba, and Kanagawa. In Japan's rental housing market, January through March is considered a peak period, especially in terms of inquiries for rental property visitations ("previews"). Therefore, this study selected two time points, March 2019 during the peak period, and June 2019, during off-peak period. For these time points, we analyzed which kind of property attributes were included in email inquiries ("responses"). To reduce heterogeneity in property attributes, floor area was limited to 15 to 30 m2 for non-married potential renters.

Descriptive statistics for the property data used in this study are reported in Table 1.

The number of properties listed within the target area in the respective one-month periods was approximately 350,000 properties for March 2019 and 160,000 for June 2019. To eliminate abnormal values, target rents were set at ¥20,000 through ¥300,000, years since construction were 30 years or less, and walking distance to the nearest train station was 15 minutes or less.

Table 1: Descriptive statistics of property data

		Rent (× ¥10,000)	Area (㎡)	Age (years)	Walking distance from railway (minutes)
March 2019	N	34,736			
	Mean	8.82	24.02	8.70	7.06
	SD	2.86	3.26	8.42	3.59
	Min	2.00	15.00	0.00	1.00
	Max	22.60	30.00	30.00	15.00
June 2019	N	15,641			
	Mean	8.80	24.27	8.60	6.57
	SD	2.75	3.23	8.00	3.37
	Min	2.50	15.00	0.00	0.00
	Max	22.60	30.00	30.00	15.00

Descriptive statistics of properties having responses (i.e., inquiries) are reported in Table 2.

Comparing Tables 1 and 2, only approximately 2% of all properties had responses, with the majority of listings having no responses[4].

Focusing on the descriptive statistics, for properties with responses, the rent was around ¥20,000 lower, and years after construction was around 1.5 years more for properties with responses. Meanwhile, no large differences (deviations) were observed in the other attributes, floor area, and distance from station. This shows that it is difficult to predict whether a specific property will generate a response based on fundamental items such as floor area and so on.

Table 2: Descriptive statistics of properties having responses

		Rent (× ¥10,000)	Area (㎡)	Age (years)	Walking distance from railway (minutes)
March 2019	N	682 (response rates : 1.96%)			
	Mean	7.02	23.93	10.28	7.19
	SD	2.23	3.66	9.12	3.58
	Min	2.60	15.01	0.00	1.00
	Max	19.80	30.00	30.00	15.00
June 2019	N	442 : (response rates : 2.82%)			
	Mean	7.90	23.73	8.31	6.99
	SD	2.45	3.60	7.97	3.53
	Min	2.58	15.65	0.00	1.00
	Max	16.40	30.00	30.00	15.00

[2] SUUMO, LIFULL HOME'S, and athome are private services that provide rental market prices usable by anyone. LIFULL HOME'S has a service called PriceMap, where a user can browse a map with pre-owned condominiums/apartments listed with estimated prices (values). TERMINAL Inc., a real-estate tech venture, offers the commercial service, Sumasate, a rent appraisal service. SRE Real Estate (formerly Sony Real Estate Corporation) also offers a price-appraisal service for pre-owned condominiums/apartments. However, none of these services has published sufficient information regarding what methods are used for what data for modelling.

[3] Listed (offered) properties are not only those of the specific company, they also include properties managed by other companies (so-called "futures properties").

[4] As for actual concluded contracts, these are determined not only by email responses by specific websites. Many other routes exist, including telephone inquiries, direct visits to agencies, referrals from other companies, and so on. Furthermore, since multiple real estate companies simultaneously list the same properties, the number of responses per property tends to become much lower.

Therefore, this study used the aforementioned three methods. For these, common explanatory variables were used in trials to predict response rates, and the respective accuracy of each method was compared.

It is generally known that there is a trade-off between model estimation accuracy and explainability. Therefore, we used logistic regression analysis as a method with high explainability, and RandomForest and XGBoost as methods with good estimation accuracy.

3.2 Variables

The explanatory variables used in this study are presented in Table 3. These include three shown in the descriptive statistics, namely, floor area, years after construction, and time to walk to station. Facilities and equipment variables, and lump sum (one-time) payments such as key money and security deposit, and so on were also added, among others. Property attributes including property type and structure were categorical values, while facilities and equipment were 0, 1 dummy variables. Dummy variables were also created from discrete variables including floor area, years after construction, and so on, and analysis was performed accordingly[5].

Moreover, deviation from market rental price was used as a special variable. This shows the deviation between estimated prices (rates) as calculated from data for analysis, and actual listed rental prices. Strictly speaking, while it is not desirable to insert estimated values into explanatory variables, in the case as in recent years where consumers can view multiple properties on Internet sites, consumers have acquired a certain sensibility such that they can recognize when a property is priced somewhat high or low. This is thought to make properties somewhat cheaper than market rental rates more susceptible to a response. Such deviation was therefore used as one explanatory variable.

3.3 Verification procedures

As there may have been a difference between responses that occurred in the peak period of March 2019 and those in the off-peak period of June 2019, estimations were made separately for data from each of these. For the verification, data were divided in an 8:2 ratio according to whether a response occurred, with 8 used as training data and 2 as data for tests. Further, since there were very few "yes" response data items, the training data was undersampled, while response probability was calculated for the test data. This trial was performed 100 times for each method, and a confusion matrix was calculated using the response probability obtained for each property. This confusion method was then used for assessments.

Table 3: List of explanatory variables

Variable attributes	Variable names
Property attributes variables	Property type
	Structure
Quantitative variales	Area
	Walk to station (time)
	Age
Facilities and euipment variables	Elevator
	Postal delivery box
	Automatic lock
	Heater toilet seat
	Floor plan
	First floor
Lump sum variables	Key money
	Security deposit
Address	City, ward, country
Custom variables	Deviation from market rent price

IV. Results

For March 2019, the distribution of estimated probability rates vis-à-vis properties which actually received responses is shown in Fig. 1.

Fig. 1 shows that, depending on the method used, there were major differences in the distribution of estimated response probabilities. For example, median values were 0.14 for the logistic regression, 0.58 for XGBoost, and 0.12 for RandomForest. Distribution shapes that differ so widely indicate that it is not appropriate to set common thresholds for response probabilities.

Therefore, in the present study, the median value of response probabilities vis-à-vis properties that actually received responses was used as the threshold. The confusion matrix was calculated as shown in Table 4, and estimation accuracy was examined.

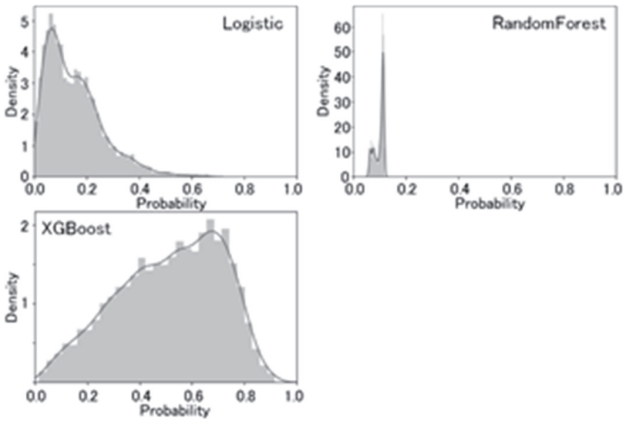

Fig. 1 Distribution of estimated response rate

[5] As for rental housing, it is known that years after construction and floor space have nonlinear effects on rent. In the analysis, dummies were created for these discrete variables, and thus they are not treated as linear.

Table 4: Confusion matrix

	Method		Estimated no response	Estimated response
March 2019	Logistic Threshold:0.13833	No response	77.00%	21.07%
		Response	0.97%	0.97%
	XGBoost Threshold:0.57917	No response	78.56%	19.51%
		Response	0.97%	0.97%
	RandomForest Threshold:0.11509	No response	67.40%	30.67%
		Response	0.97%	0.97%
June 2019	Logistic Threshold:0.13114	No response	74.63%	22.56%
		Response	1.41%	1.41%
	XGBoost Threshold:0.536389	No response	73.37%	23.81%
		Response	1.41%	1.41%
	RandomForest Threshold:0.10777	No response	63.36%	33.83%
		Response	1.41%	1.41%

Here, when the focus is on the detection rate where there was no response both for actual properties and in the estimates[6], it can be seen that these were roughly the same for the logistic regression and for XGBoost. While XGBoost shows a higher detection rate for March 2019, the logistic regression shows a high detection rate for June 2019. In reality, the peak period of March 2019 was of greater importance, as there were more listed properties and more concluded contracts. In this sense, it can be argued that XGBoost had a somewhat higher accuracy than the logistic regression.

In general, in classification problems of this nature, it is known that RandomForest and XGBoost are able to make classifications of high accuracy. However, when thresholds are adjusted, the traditional logistic regression method can be considered to show sufficient classification accuracy for practical purposes.

V. Considerations and future issues

The present study compared the results of three different methods: logistic regression, XGBoost, and RandomForest. In this case, the use of classic logistic regression was considered desirable for actual business applications.

When logistic regression and XGBoost have mostly the same prediction accuracies, logistic regression results are easier to interpret. With logistic regression, it is easier for businesspersons to understand reasons for using estimation results to make decisions.

Certainly, there are some cases even in business when, in the use of statistical methods, accuracy takes priority when there is no need for explanatory power or interpretability. However, if accuracy rates are roughly equivalent, the logistic method is easier to interpret and understand. In business sites where the statistical results are to be used, it is easier to form a consensus using logistic results. This in turn reinforces its usage in business and is also thought to be connected with model refinement.

When engaging in model development, it is essential to have a grasp of the latest statistical methods, so as to be able to use them when needed. This does not mean that use of the latest methods is necessarily the best choice. For business applications, multiple methods should be tested and the statistical method to be used should be selected while considering the proper balance between accuracy rate and explanatory power (interpretability). One must also always be aware of the risk of overfitting, depending on statistical method used.

A future issue is investigation of what differences occur in results depending on the statistical method used after data have been appropriately stratified. One assumes that different stratification approaches have differing results. Thus, there may be a possibility that the appropriate stratification method may depend on what statistical methods are to be used.

REFERENCES

[1] H. Watanabe, Y. Ichifuji, M. Suzuki and S. Yamashita, "Statistical modeling of transition time until occupation of rental rooms using housing information website data", The Japan Society for Artificial Intelligence, JSAI2019(0), 1D3OS10b03-1D3OS10b03, 2019.(in Japanese)

[2] Y. Aota and Y. Tanaka, "Analysis of factors of price determination and price difference within online transactions: Focus on hotel room prices using a hedonic approach", The Japan Society of Household Economics, Vol 22.23, 71-79, 2006.(in Japanese)

[3] F. Yamazaki, Y. Asada, H. Seshimo and C. Shimizu, "Empirical study of tenure choice and user cost for housing", Journal of the Housing Research Foundation "Jusoken," 33(0), 335-345, 2007.(in Japanese)

[4] T. So and Y. Arai, "The influence of the accumulation of wealthy people and public husing estate clusters on rent", Urban housing sciences 2018(183), pp.126-131, 2018

[5] T. So, "Effect on the price of pre-owned condominiums of apartment vacancy rates within a locality", Journal of the Japan Association for Real Estate Sciences, 32(1), 106-113, 2018.(in Japan)

[6] T. So and C. Shimizu, "Housing Facilities and Housing Rent, Purchasing and Supply Management", (Peer review book), ISBN 978-1-78984-973-8, 2019.

[7] C. Shimizu. "Methods of real estate price determination from a big data perspective", Journal of the Japan Association for Real Estate Sciences, 31(1), 45-51, 2017.(in Japanese)

[8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system", arXiv:1603.02754, 2016.

[9] L. Breiman, "Random Forest", Machine Learning 45, pp.5-32, 2001.

[6] In actual business, business efficiency is improved when emphasis is placed on selling properties having a high likelihood of response. The merits of this approach are therefore high.