Statistical Methods for Machine Learning

Winter Term 2021/22

Institute of Informatik

M Sc. Informatik AFB 2020-Participants

Athira Mavoomkuttathil Sivachandran - 526586

Nisha Muthuraju - 521440

Srikanth Gelli - 524656

Vinendhar Reddy Bollu - 526328

TECHINICAL UNIVERSITY OF CLAUSTHAL

Dr.Marius Ötting

14-01-2022

Table of Contents

1.Introduction:

As money being a crucial factor for renting out one's room or an apartment it does help the hosts to find the right price per night for their property. Most of the hosts are not business people, but an owner of a personal apartment who lack the domain specific knowledge of running a hotel. Here comes the statistical learning methods for the rescue. These methods can provide some insights based on historical data and can help hosts make some profitable decisions.

When it comes to price per night, host just cannot give away the property on lease for a cheap price not make the price exorbitant merely based on his/her gut feeling because it can be extremely subjective and the housing market being very competitive some other host might beat them to it with the right price. But it should be optimal, in order to make that decision host should understand what factors drives the customer's decision and what factors should be considered for finalizing the prices. With the help of real-time data (provided it is available) host can also make a dynamic pricing model to find a balance between profit and loss. Thus, hosts can make an optimal and profitable pricing strategy based on statistical learning.

As observing the variables and running some codes in r studio, the number of nights the customer stayed is highly related to the price of the apartment or room. Moreover, there are quite other number of variables in the dataset that determine the price factor of the room, such as the number of rooms in a single apartment, response and review of the experienced customers and many more, which are interesting to investigate by doing this project using R programming language.

2.Exploratory data analysis

For analysing the data, the given data set is loaded into R studio using read command. For the initial observation the number of columns, rows and their total and number of missing values are derived.

The provided dataset contains 3000 rows, which indicate apartments in the city called Boston against 33 columns of different variables that designed the data frame. Yes, there are missing prices. For few variables, mean value of the column is calculated and attributed to the missing places with this mean value. For some variables we have decided to omit/delete the missing values (NA values)

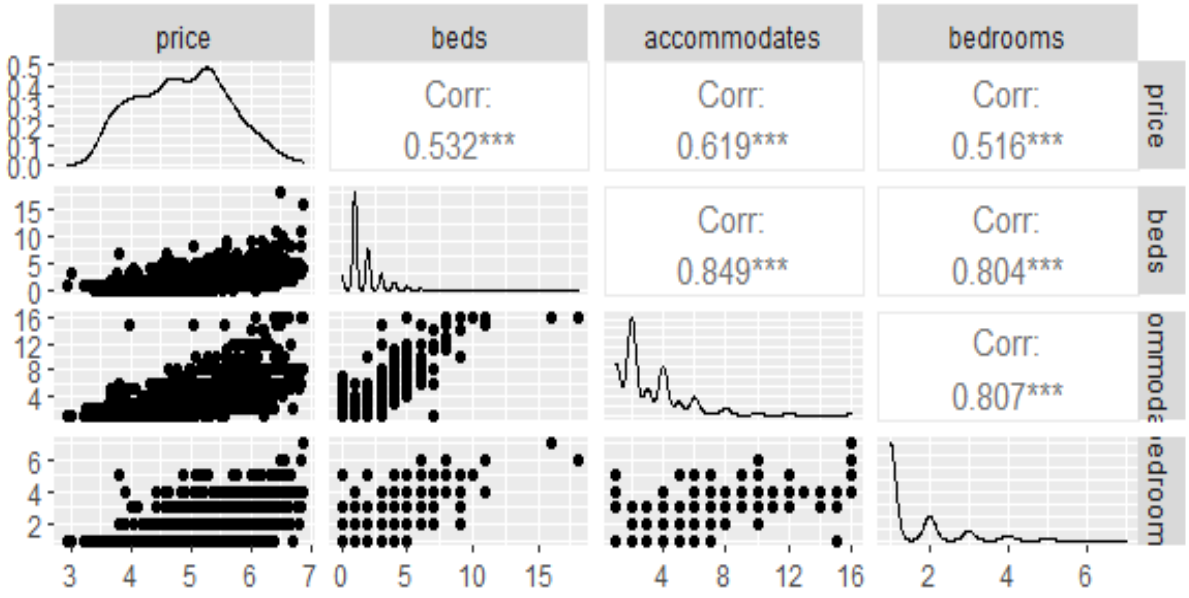
We have tried to avoid the variables where there are high number of missing values. Below are the three variables which are likely to affect the price. We have also calculated the Pearson correlation coefficient between the variables and the prices.

1.beds - 0.532

2.bedrooms - 0.516

3.accommodates - 0.619

So, the beds, bedrooms and accommodates will affect the price.



From the above graph we have analysed 3 highly correlated covariates. Beds and accommodates 84.9% correlated and beds and bedrooms 80% and accommodates and beds 80 %.

3.Investigate drivers of prices per night

To investigate the drivers of price per night multiple linear regression is used.

For the purpose of prediction of a response variable (here it is 'price') we need a model. In general, the simple linear regression is used to predict a quantitative outcome y on the basis of one single predictor variable x. The goal is to build a mathematical model (or formula) that defines y as a function of the x variable. Once, we built a statistically significant model, it's possible to use it for predicting future outcome based on new x values.

For the given dataset we have multiple distinct predictor variables and so we considered Multiple linear regression model.

Multiple linear regression is an extension of simple linear regression used to predict an outcome variable (y) based on multiple distinct predictor variables (x). With three predictor variables (x), the prediction of y is expressed by the following equation:

$$y = b0 + b1*x1 + b2*x2 + b3*x3 + \varepsilon\ldots\ldots\ldots \text{ (equation 1)}$$

The "b" values are called the regression weights (or beta coefficients). They measure the association between the predictor variable and the outcome. "b_j" can be interpreted as the average effect on y of a one unit increase in "x_j", holding all other predictors fixed.

If the Y value is non-negative this might lead to non-sensical, negative predictions.

So we can transform the Y* as log(Y) and use the multiple linear regression model for Y*

$$Y^*\, i = \beta_0 + \beta_1\, x_{i1} + \beta_2\, x_{i2} + \varepsilon_i \ldots\ldots\ldots\ldots \text{ (equation 2)}$$

And here in this project log of the response variable is taken.

We are fitting the model by "lm()" function in R studio we get the intercepts ,and beta values.

```
lm(formula = price ~ ., data = Airbnb_5)

Residuals:
     Min       1Q   Median       3Q      Max
-1.51359 -0.30869 -0.01944  0.26053  2.16781

Coefficients: (6 not defined because of singularities)
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            4.724e+00  2.060e-01  22.930  < 2e-16
host_response_rate                    -7.797e-03  1.522e-03  -5.122 3.27e-07
host_acceptance_rate                   1.448e-04  5.766e-04   0.251 0.801738
host_listings_count                    3.222e-04  3.453e-05   9.332  < 2e-16
host_total_listings_count                     NA         NA      NA       NA
accommodates                           8.954e-02  9.234e-03   9.697  < 2e-16
bedrooms                               6.851e-02  2.059e-02   3.327 0.000892
beds                                   2.301e-03  1.430e-02   0.161 0.872195
minimum_nights                        -2.162e-03  2.383e-04  -9.074  < 2e-16
maximum_nights                        -2.027e-05  2.036e-05  -0.996 0.319463
availability_365                       1.053e-04  7.684e-05   1.370 0.170845
number_of_reviews                     -1.854e-04  1.523e-04  -1.217 0.223704
number_of_reviews_ltm                  3.027e-03  9.407e-04   3.218 0.001310
number_of_reviews_l30d                 9.719e-04  6.253e-03   0.155 0.876492
calculated_host_listings_count        -2.355e-03  3.160e-04  -7.454 1.25e-13
days_as_host                          -4.004e-05  1.021e-05  -3.922 9.04e-05
`host_response_time_a few days or more` -7.482e-01  1.409e-01  -5.311 1.19e-07
`host_response_time_within a day`      -1.299e-01  4.614e-02  -2.816 0.004904
`host_response_time_within a few hours`  2.513e-02  3.875e-02   0.648 0.516814
`host_response_time_within an hour`    3.994e-03  2.596e-02   0.154 0.877771
host_is_superhost_f                    1.272e-01  2.237e-02   5.685 1.47e-08
host_is_superhost_t                           NA         NA      NA       NA
host_has_profile_pic_f                 2.628e-01  3.262e-01   0.806 0.420501
host_has_profile_pic_t                        NA         NA      NA       NA
host_identity_verified_f              -6.236e-02  2.651e-02  -2.353 0.018713
host_identity_verified_t                      NA         NA      NA       NA
Entire_home_apt                        9.071e-01  1.353e-01   6.704 2.51e-11
Hotel_room                             1.330e+00  1.763e-01   7.544 6.42e-14
Private_room                           2.298e-01  1.346e-01   1.707 0.087876
shared_room                                   NA         NA      NA       NA
instant_bookable_f                    -3.312e-02  2.241e-02  -1.478 0.139531
instant_bookable_t                            NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4593 on 2433 degrees of freedom
Multiple R-squared:  0.645,     Adjusted R-squared:  0.6414
F-statistic: 176.8 on 25 and 2433 DF,  p-value: < 2.2e-16
```

Fig 1

By observing the p values. if $p < 0.05$ then they are statistically significant variable,they affect the nightly rent price. For example, 'Accommodates' and the other variable is 'bedrooms' with $p < 0.05$. Some of the significant variables are" number_of_reviews_ltm ","host_response_time_within a day`","host_identity_verified_f", etc.

If the regression coefficient is increased by 1 unit(accommodates) this will affect the price by 1.2 times and value of p is statistically significant. Similarly other variables p value can be observed and conclude whether they are statistically significant. From the above figure we can see the multiple R ^2 value is 0.645 out fitted model will explain this 64.5% of data.

log(price) = 4.72e+00 + -7.7e-03*host_response_rate + 8.954e-02*accommodates + ….

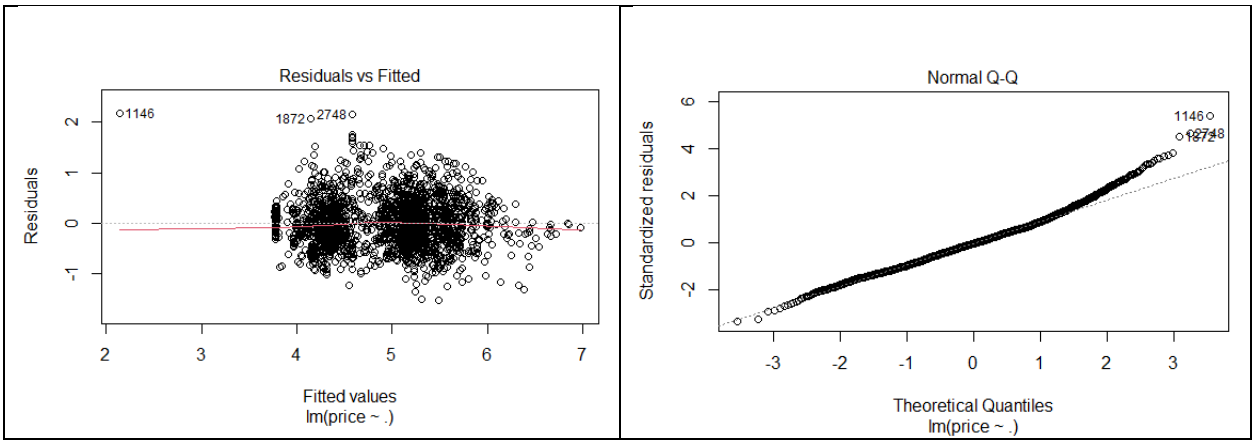
By analyzing this linear model, the estimated coefficients of bedrooms and accommodates are used to predict the nightly rent price using the data 'Airbnd_5'(variable holding the transformed data from original data)

We applied a linear model using the function lm() the r code is..

lm(formula = price ~. , data = Airbnb_5)

Summary(model2)

Plot(model2)



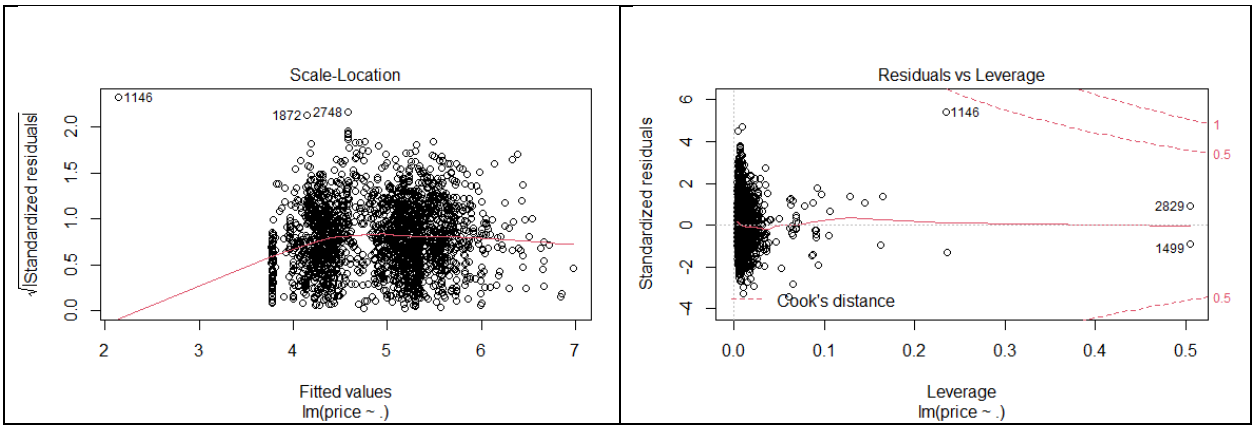


Fig2

Figure-1(top right counted clock wise) Residual vs Fitted – we plot the estimated response values on X-axis and residual on Y-axis . Data from transformed dataset is highly concentrated around the central line and hence it has approximately linear. So there is a linear relation between response values and residual values. From this we conclude that model is fitted for the data. From normal QQ plot says the normally distributed. Picture 3 There is a constant variance along the regression line and it is a good model to predict price.

4.Comaprison and classification Different Methods

The packages glment () was used and the dataset Airbnb _5 is spit into 75 % and 25 % respectively and assigned to an index, that is ‘x_train’ in R sheet attached, the rest of the data that is test data set is obtained by adding a minus sign to the index variable. The testing_data is the data frame with 1844 observations and the test_data with 615 observations. The data set is spit into training set and test set because, it is not need to train the entire data set for making prediction. Here the goal is to predict the nightly rental price in regards to all other variables (covariates) and evaluate how it performs. The set.seed() random function is used for getting same results again and again while testing the code.

Here the incorporate two other models other than multiple linear regression and test the data model to predict which model is better to predict the price.

Appling the same linear regression model in this section at first, but now fitting the model to the training data. the trained model is used to predict the prices of the test data (other 25 %)

In the previous exercise Multiple linear regression is use and, here for training and testing the data other two models are using one is Ridge regression and Lasso regression.

In ridge regression is a shrinkage method which uses least square to fit a linear model, as the ridge coefficients approaches towards zero this technique is best for fitting the model. Ridge regression have slightly better than other regression model. In Ridge regression looks the coefficient estimates that fit the data well by making RSS small.

$$-\ell(\beta) + \lambda \sum_{i=1}^{p} \beta_j^2,$$

(Equation 3)

Equation for ridge regression

Other than Ridge regression we use Lasso model, to take the variables exactly, not like in Ridge regression Ridge regression the model, may take all predictors values which may leads to wrong interpretation. The Lasso regression overcomes this even thou they are similar in formulations. The models obtaining after the evaluation of lasso it force some coefficient to be zero when the lambda is large and ensure variable selction.so the model generate ed after lasso will be easy to interpret.

$$-\ell(\beta) + \lambda \sum_{i=1}^{p} |\beta_j|,$$

(Equation 4)

Equation for Lasso

In Ridge regression the lamda, shrinkage penalty parameter is used to avoid the multiple highly corelated variables, that is if there are lot of variables which are highly corelated to each other, it will help to reduce correlation.

“coef” is taken in both models that is to obtain coefficient of variables. By taking the coefficient of variables, initially obtain an intercept. The highly significant variables are bed, bedroom and accommodates have better effect on the predicted price.

The predicted price is then compared with the actual price that is the values in "test_data" columns is compared with "ridge. predit. bestlam". The Ridge regression have no much higher difference in the prices which predicted to the actual value. Hence by doing this project it was easy to arrive an conclusion that ridge regression method is slightly better than other models.

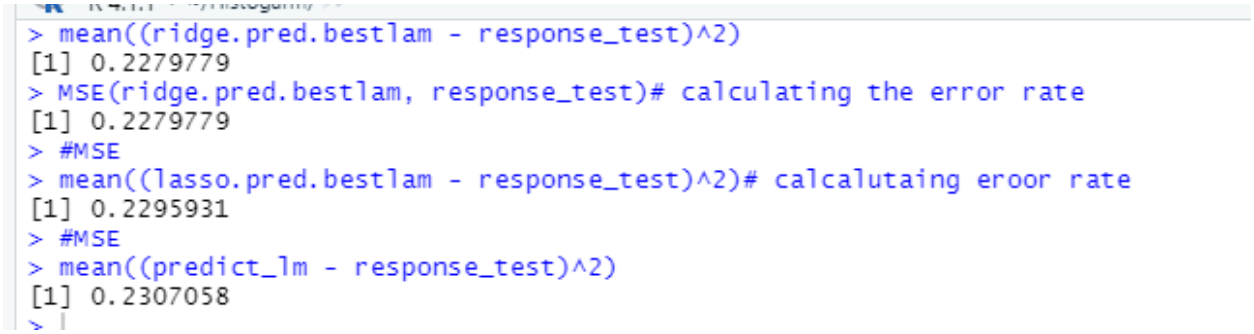


Fig3

The MSE of all three models used in this project is calculated for analyzing which model is better. As the interpretation from the above r code a conclusion is made that ridge regression have lesser error rate so that the model fit correctly. The MSE value of Ridge regression is 0.2279 which is the smallest error rate out of other models mentioned in this project.

The lamda values for the Ridge regression and Lasso is as 0.053 and 0.0024 respectively. And the beta values can be obtained from adding intercept with coefficient with corresponding x values (covariates).

The final response value that is price is log value by calculation so by taking the anti-log of the values of reasponse value, the actual rates can be obtained.

Conclusion

Implementing this project in r studio with the provided data set was quite challenging for the beginners of R language. However, it was interesting and easy to learn the concepts with guide of lecture slides and other tools. By investigating the data frame, we come up with two best models to interpret the data and predict the price, they are multiple linear regression and Ridge regression. Even though the Ridge regression take all variables for calculating the coefficient which may lead to mis interpret the model, the error rate is considerably less and can be fitted to model the data set provided. The log of response variables is taken in our case, and so the genialized equation becomes log (y) become Y*.

$$\log(y) = Y^* = b0 + b1*x1 + b2*x2 + b3*x3 + \varepsilon \text{ (equation 5)}$$

By investigating the data Ridge regressions model can be found to be the best model to finding the nightly rent price. The Car parking area, carpet area (outside the room for keeping their foot wears and other stuffs), Pets allowed, Kids rooms, Kitchen availability (which were not included in the room type) and garden or playing area would be enhance the rating and review of the apartments or rooms and can make the data set more precise like city like Boston. Moreover, it may be affecting the data analysis which have been done yet, the covariates related to host acceptance can be more precise. There were a large number of missing values in covariates so it was tough for interpreting, hence deleted. Due to pandemic online group organization and communication was weak and hence affected the progress of the report was an remarkable problem which faced.