Talk a little bit about myself

- I'm a second year master student in Economics
- My background is in finance and banking. I spent 2 years working for a commercial bank in Vietnam.
- My interest in applying statistical learning in forecasting started when I took a course on econometrics. I worked on a project on forecasting the electricity price using time series data. Since then, I have focused more on data mining, machine learning, and data science. My favorite book is "An Introduction to Statistical Learning" by Gareth James, Trevor Hastie. Besides reading, I took some courses on data analysis, applied machine learning, and programming.

Feel free to check out my Linkedin: https://www.linkedin.com/in/tuong-andy-vu-312222182/

- I would consider myself more practical application-oriented.

- I'm familiar with R, Python, SQL, and Java.

- Now I would like to share with you what I want in my career path :

-
    - I would be thrilled to collaborate with you on potential research topics on finance or economics.
        - After graduating, we plan to take an internship or fellowship in the related fields.

Motivation & Objectives

Hello. So the purpose of my thesis was to find a way to predict the Airbnb rental price in New York City. Let's say you are an Airbnb host, you want to maximize your profit. How can you do that?

You have to learn how to set the price in a smart way?

But it's not easy for an ordinary host.

So the aim of my project is try to answer two big questtion?

1. Which is the best model to predict the an Airbnb rental price given its characteristic such as number of rooms, bathrooms it has, the review rating score, the neighbourhood it is in . . .
2. An other important question is that among many features, which are essential in predicting the rental price?

The Workflow

Before going into details, I want to show you my workflow to answer those two question

We collect the data then we prapare data to make it ready for modelling and then we use several models to fit the data, finally we evaluate the prediction performance of each model.

Turning to our problem

First let's take a look at our study area

We can see the distribution of Airbnb listing in New York City. We perform several procedures in the data preparation step, such as handling missing data, filtering data, transform data . . .

The final data set contains about 40000 observations with a huge number of features.

We then visual the data to explore the relationship between the rental price and each feature. For example we can see that

A listing's price seems to be positive associated with the number of guests it can accommodate.

And

The location might have an important role in determining the price. Unsurprisingly, listing in the city center (Manhattan and Brooklyn) has a higher median rental price.

And

Hosts with superhost status usually charge higher prices. A possible explanation for this might be thatpeople are willing to pay a premium price because they consider superhost status a mark of quality.

And

Hosts with verified profiles gain a price premium. The relationship may be explained by the fact that verified profiles (e.g., by providing ID and verifying your phone number and email address) can increase their trustworthiness and, therefore, can charge a higher rental price.

After getting a sense of our data we turn to the main question : How to predict the price of an particular Airbnb listing given the information about it.

Well, we build model!

We have many models that can be used to predict the price. But how can we tell one model is better than others?

We can assess how well a model performs by using mean square error criteria:

The best model is the one that has the lowest test MSE (not training MSE)

Turn out, the MSE can be decomposed into three parts :

The question is *is there a model that have lowest bias and lowest variance?*

It's seem impossible to have a model like that. . Too simple model (low variance) tends to have high bias and too complex model (low bias) is likely to vary significantly. This is known as Variance-Bias tradeoff or variance-bias dilema in statistics and machine learning.

So, How can we deal with that dilemma, what is the solution?

Well, we just try several commonly-used algorithms to find the one with the lowest test mean square error:

Linear Regression

Ridge Regression

Lasso Regression

XGBoost

We got the result in the table in slide 15

. . .

We still have another question to answer: Which features are important to predict the rental price?

XGBoost or any decision-tree based methods has a useful feature of identifying important features.

We list top ten important features here. . .