Questions

1. Please describe the central limit theorem and provide an example.

Given a series of independent and identically distributed random variables with sample mean and sample variance, bound to under infinity, the central limit theorem states that the distribution of the sample means will be approximately normally distributed given enough samples, the sample mean will converge towards the true population mean as , and the variance of the distribution of the sample means will shrink if you have more samples.

An example can be let’s take

1. Describe a classification algorithm that you have previously put into production and why it was chosen.

1. Describe the difference between bagging and boosting methods, and when to use one or the other.

Bagging methods use many weak classifiers trained on different features of the data chosen randomly without replacement. The average of all of the predictions (for regression) or the majority vote (for classification) becomes the final prediction. A good example of a bagging method is Random Forest.

Boosting methods create a collection of weak classifiers, where each classifier learns from the errors of the prior classifier. The final classifier would then be the best perofrming model.

It has been shown that boosting generally performs better than bagging.

1. Describe 2 regularization techniques for a random forest model

Data Challenge

Attached you will find the file listings.csv which provides data on around 50,000 AirBnB listings in New York City.  Given only this data, you want to create a model to predict how much you can charge for new listings while keeping vacancy down.

Please put together a brief analysis of the dataset and show how you would go about creating a model to predict a listing price, while taking market demand into account.