LINEAR REGRESSION VS RANDOM FOREST REGRESSION FOR AIRBNB BOOKING PREDICTION MODEL

Let me first explain the differences between linear and random forest regression before diving into which one is a better use case for bookings.

The key concepts of random forests are, random sampling of training observations when building trees and random subsets of features for splitting nodes. Random forest regressions also discretize continuous variables since they are based on decision trees, which function through recursive binary partitioning at the nodes. This effectively means that we can split not only categorical variables, but also split continuous variables. Additionally, with enough data and sufficient splits, a step function with many small steps can approximate a smooth function for predicting an output.

A Linear Regression model, just like the name suggests, created a linear model on the data. A simple way to think about it is in the form of y = mx+C. Therefore, since it fits a linear model, it is able to obtain values outside the training set during prediction. It is able to extrapolate based on the data.

Now let’s see how each model is applicable to Airbnb’s bookings. We need to understand what features exist in our dataset. We can assume the dataset will have features like:

location features

Seasonality

number of bedrooms and bathrooms

private room, shared, entire home, etc...

If we were to measure the price of bookings in just one city, we could probably fit a decent linear regression.

Take Nairobi for an example, the coefficient for each bedroom, bathroom, time of month, etc… could be standardized across the city if we had a good variable that could take into account location in the city.

Given the effect of different events that influence pricing, we could create custom interaction effects between the features if, for example, a huge conference suddenly increases the demand of three or four-bedroom houses.

However, let’s say we have thousands of features in our dataset to try and predict prices for different types of homes across the world. If we run a random forest regression model, the advantages are now forming complex non-linear combinations into a model from a dataset that could hold one-bedrooms in Nairobi and mansions in Dubai.

But if our problem set is back to a simple example of one zip code of Nairobi, then our feature set is dramatically reduced by variation in geography and type of rental, and a regular linear regression has benefits in being able to understand the interpretability of the model to quantify the pricing factors.

Given this bit of information my answer is for a specific zip code e.g. Nairobi Airbnb should use linear regression to build a model to predict booking prices but for a larger geographic area random forest test regression should be used to cater for the vastly increased variables.