Take Home Problem: Data Scientist

The questions below are meant to give candidates a sense of the problems that we tackle at Opendoor. We expect solutions in the form of a readme + working code. Please limit yourself to 3 hours to complete this problem set.

Problem: *k*-NN strikes back!

Let's look at a collection of home sales available here. In this problem, we will build a simple pricing model using k-nearest neighbors.

k-NN formulation

Suppose

- p_i is the close price of home *i*
- N_i is the k-nearest neighbors of home *i* by spatial distance

The *k*-NN model predicts the price of home *i* as

$$\hat{p_i} = \sum_{j \in N_i} w_j p_j$$

where w_j is the weight given to p_j (the close price of home *j*) and should sum to 1.

Questions

- Using the dataset provided, please build a *k*-NN model for *k = 4* that avoids time leakage (details below).
- What is the performance of the model measured in Median Relative Absolute Error?
- What would be an appropriate methodology to determine the optimal *k*?
- Do you notice any spatial or temporal trends in error?
- How would you improve this model?
- How would you productionize this model?

Implementation notes

- To prevent time leakage, a home *j* should be considered a neighbor to home *i* only if the close date of *j* occurred prior to the close date of *i.* Think about making a prediction using information available to house *i*. You only want to use information you have available at that time. One way of doing this is to restrict yourself to neighbors that have closed prior to the close date of *i*.
- The Median Relative Absolute Error (MRAE) is defined as $median\left(\frac{|\hat{p_i} - p_i|}{p_i}\right)$