

TECH ALPHARETTA

2972 Webb Bridge Rd, Alpharetta, GA 30009

**Project # 2 : Property Click Prediction**

*This data project has been used in the recruitment process for the data science positions at NoBroker Data Sciences.*

Properties form one of the most important data entity in nobroker data ecosystem. Properties receive interactions on NoBroker. One interaction is defined as one user requesting an owner contact on a property. A property can receive 0 to many interactions. This research will focus on studying and modelling the interactions received by properties.

**Assignment**

We are interested in studying and statistically modelling property interactions. **We would like to have a predictive model that would say the number of interactions that a property would receive in a period.** For simplicity let’s say we would like to predict the number of interactions that a property would receive within 3 days of its activation and 7 days of its activation. However, this part is open ended and you could bring your own time intervals into the problem. This is the part of your artistry in data science. In the end we need to profess the number of interactions that a certain kind of property would receive within a given number of days. We cannot do a time series forecasting here considering the limited amount of data that could share as a part of an assignment. You may clean the data, merge them, do an EDA, visualize, and build your model.

**Data Description**

Unzip the datasets.zip file to find the following 3 data sets:

property\_data\_set.csv :

* Properties data containing various features like activation\_date, BHK type, locality, property size, property age, rent, apartment type etc.
* activation\_date is the date property got activated on NoBroker. Fields like lift, gym etc are binary valued - 1 indicating presence and 0 indicating absence. All other fields are self-explanatory.
* You may use these along with the rest of the data sets to engineer the features that you would use in your study.

property\_photos.tsv :

* Data containing photo counts of properties.
* photo\_urls column contains string values that you have to parse to obtain the number of photos uploaded on a property.
* Each value in the photo\_url column is supposed to be a string representation of an array of json [ in python terms a list of dictionaries] where each json object represents one image. However due to some unforeseen events, these values got corrupted and lost their valid json array representation. You could see this if you observe the data closely. Hint: There is a missing “before ‘title’ for the first json object in each value. There is also an additional “at the end of each value. Also, you must remove all the \\ to get a valid json representation.
* Your objective is to get the number of photos uploaded for a property. For this you should correct the corrupt string and make it a valid json. Once you have a valid json string, you can get the length of this array, which would be the number of photos uploaded on the property.
* Also note that these are not images, but just names that we use to point to images. You are NOT given the images, nor do we expect you to have them. All that you are expected to do it get the number of photos on each property by cleaning up the corrupt invalid json array string.
* NULL/NaN values indicate absence of photos on the property, ie; photo\_count = 0

property\_interactions.csv :

* Data containing the timestamps of interaction on the properties.
* Each request\_date value represents the timestamp of a unique valid interaction on a property (contact owner happened and a user received the owner contact phone number)
* Therefore, if you count the number of times each property has appeared in this table, it tells you the number of interactions received on this property
* You will use this request\_date along with the activation\_date in our first table and other features in our study.

**Practicalities**

Please go through all the instructions and data descriptions carefully before getting on the ground.

We DO NOT look just at your final model and its performance, rather we look for the research mindset in you, your curiosity in data, your enthusiasm to collaborate and if your work mindset fit in our DS culture. Therefore, we urge you to present whatever you do with standards followed among the data science community. Keep an open-ended eye on the problem and feel free to approach the data in whatever way you think suits the problem. We urge you to try out different methodologies and present your results.

This is a research assignment project and please don’t struggle for a full-fledged hurried submission. Quality is what We believe in, and We also believe Great things are built small bit at a time. Hence present everything and anything you have done and strive for good quality in them.