NoBroker Data Sciences

Project Falcon

Data Sciences Team Compatibility Test

April 2017

Instructions:

- Please go through the entire document carefully before getting on the ground
- You may take 72 hours or less on this problem and submit whatever progress you have made. This is a research assignment 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.
- If you need any help or clarification or extra information, please feel free to contact us. We think like researchers and we believe in collaboration. We are not invigilators nor your bosses. Please feel free to ask for help. Knowledge is Free.

Introduction

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.

Problem Statement

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 of time. 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 interaction 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 shared as a part of an assignment. You may clean the data, merge them, do an EDA, visualize and build your model.

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.

Data Sets

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

a. **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

b. **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

c. **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 interaction 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

Note: property_id is the unique property identifier in all the three data sets.

Tech Stack

At NoBroker, we are building the best in class Data Science, Machine Learning & Artificial Intelligence platform for bringing real value to our customers. Therefore it would be very much desirable to have the following tech tools in your submission.

1. We have adopted python for deploying scalable, real time, consumer facing machine learning models. Our ML stack is fully in pure python. But we do use R for BI data science.

2. Therefore, it is required that you do at least the modelling part in one of the of the Python's machine learning modules. You might use Scikit Learn, Tensorflow or any other Pythons ML packages. We believe Python is very sweet just like our people.

3. You may use R or Python, whatever your choice is, for doing the data cleaning and aggregation part. If you are using R, please make sure you only prepare the data in R. Whatever model you are building is desired to be built in Python. Therefore if you prefer R, you could prepare your final-model-ready csv file in R and then load it using pandas into Python and subsequently plug into your Python's ML module

4. It would be nice, if the submission happens in the standards of collaborative research in the Data Science community. For R you may RMarkdown and for Python - Jupyter Notebooks. These are desirable but not necessary. What is very necessary is a quality documented code.

Guidelines

1. You may start with cleaning the property_photos data set and getting the number of photos on a property.
2. You may merge the data sets and then make appropriate transformations to our data set to be ready for any EDA that you would be interested in.
3. You could play with your data, make visualizations, pose questions, and understand the challenges.
4. You may then do a feature engineering for your modelling part and get your data ready for training.
5. You may try upon a few models and see if there is any of your hypothesis can be proven
6. You will benchmark your model with the metrics best suited for your algorithm. For example: accuracy, RMSE etc.
7. Please present whatever you do. From beginning to whatever you have done. This is not a graded test. We want to see your mindset not the solution

Submission

1. If you are using Jupyter Notebooks or RMarkdown or both together. Please submit them all
2. Please put all the code and data you have used in the submission. This has to be a reproducible research
3. Please document your code and ensure good quality. Nothing is more satisfying than reading a quality code.

You might want to try

The following part is NOT AT ALL expected in the submission, but is put here because we are just very enthusiastic people ;)

1. You might want to try building a simple REST api to serve your ML model. Tips: Use Python's Bottle or Flask micro framework for minimal effort
2. You might want to use Docker to pack up all the dependencies and code into a single image so that anybody could run it right out of the box. If you have never tried docker, please don't even think of learning it in 72 hours

Thank you very much
Have much Fun