Real Estate: Lead Conversion

Problem Statement

Develop a scoring system for lead conversion and derive online behavior analysis along with recommendations to client.

Data

There are two data files - one containing the information on online beavior of the leads and the other containing information of those leads being converted to tenants. These two files, henceforth will be addressed as `*leads_data*` and `*taregt_data*` and the target column `converted_to_tenant` as `*output*`.

Approach

The primary tasks to accomplish the goal of this project are:

- Understand the data
- Map `target_data` to `leads_data`
- Develop a scoring mechanism

The solution follows a standard Data Science solution approach targeting the following sections:

Insights into the Data

- Some lead ID values have multiple outputs in `target_data`
- There are multiple data points repeating information in `target_data`
- `lead_id` of `target_data` does not have the same column name in `leads_data`
- There are no common IDs in `target_data` and `client_id`, hence use of `ga_lead_id` to merge with the `target_data`
- Number of leads are more than number of clients, hence a client might be associated with more than one lead ID
- Grouped columns based on their characteristics and identified which to drop and those which need advanced analysis
- Generated client level and leads level numeric, categorical and boolean features based on the above grouping

Hypothesis 1

A client has multiple interactions with the website – positive result; as there exists duplicate data points for a client but we do not have enough evidence to drop them.

Example: `client_id` = 4855140000000000000 has 34 out of 49 unique interactions and thus a 30.6% of duplicates.

Hypothesis 2

A client might be linked with multiple leads – positive result; as 24.6% clients have been assigned more than 1 lead ID.

Hypothesis 3

Each `lead_id` is associated with only one client – negative result; as there are 3 client IDs associated with more than 1 lead ID which constitutes 0.07% of the data.

Hypothesis 4

The number of data points for a particular `client_id` affects the value of `output` - positive result; as on checking importance of features the model is trained on, higher importance is given to the features generated on client level using the feature design mentioned in the notebook.

Insights from Model

- Built a scoring mechanism that will help the sales executives prioritize the leads using Random Forest Model
- Originally the data had a 17% success rate of converting leads to tenants
- Using this model we can achieve a 19% increase, i.e a total of 36% success rate in lead to tenant conversion while selecting top 50 leads

Recommendations to Client

Recording more information on the following will help online behavior analysis and improve scoring model

- Number of pages per session directly affects the conversion rate
- Number of sessions directly affects the conversion rate
- Bounce rate inversely affects the conversion rate
- Some categories of `Source` directly affects the conversion rate
- Lead ID assignment methodology - understand and refine the logic for assigning lead IDs eg: a lead ID is assigned to only some rows of the same client even though the others contain same information