Santader Product recomendation

Problem :
Build a supervised learning recommender system given a little over 1.3M denormalized dataset of customer details and associated products.

Dataset:
Training set: 1.3M
Total 48 columns : First 22 columns represent customer information comprising for both categorical and continuous variables. The remaining columns ending with “ult1” are products that the customers have.

```
[u'fecha_dato', u'ncodpers', u'ind_empleado', u'pais_residencia',
       u'sexo', u'age', u'fecha_alta', u'ind_nuevo', u'antiguedad', u'indrel',
       u'indrel_1mes', u'tiprel_1mes', u'indresi', u'indext', u'canal_entrada',
       u'indfall', u'tipodom', u'cod_prov', u'nomprov',
       u'ind_actividad_cliente', u'renta', u'segmento', u'ind_ahor_fin_ult1',
       u'ind_aval_fin_ult1', u'ind_cco_fin_ult1', u'ind_cder_fin_ult1',
       u'ind_cno_fin_ult1', u'ind_ctju_fin_ult1', u'ind_ctma_fin_ult1',
       u'ind_ctop_fin_ult1', u'ind_ctpp_fin_ult1', u'ind_deco_fin_ult1',
       u'ind_deme_fin_ult1', u'ind_dela_fin_ult1', u'ind_ecue_fin_ult1',
       u'ind_fond_fin_ult1', u'ind_hip_fin_ult1', u'ind_plan_fin_ult1',
       u'ind_pres_fin_ult1', u'ind_reca_fin_ult1', u'ind_tjcr_fin_ult1',
       u'ind_valo_fin_ult1', u'ind_viv_fin_ult1', u'ind_nomina_ult1',
       u'ind_nom_pens_ult1', u'ind_recibo_ult1']
```

Prediction : Predict additional products customers will get in the last month ie 2016-06-28.

Solution Method:

1. Exploratory data analysis and cleaning
2. Model
3. Parameter tuning
4. Evaluation

1. Exploratory data analysis

Basic exploration and cleansing of raw data set was done. Related Notebook is attached.

2. Model:

The data is almost 50% divided between predictor variables and target variables. After cleaning for null and missing values we still have 1.3M rows, across 929k customers. A decision tree based solution is preferable given the low dimensionality of the predictor variables. Xgboost library was used as it is simpler to implement and usually performs very well with out of the box variables. Model is fit using 5 fold cross validation.

3. Parameter tuning

Code for exhaustive grid search for hyper parameters, but due to lack for compute was not run to full completion.