CROSS SELLING RECOMMENDATION

Group Name

LISP01-Data-Analysts

Name, Email, Country, College/Company, Specialization

Lauro Ribeiro, lauroc.r.volei@hotmail.com, Ireland, Data Glacier, Data Analyst

Lasisi Salmah, lasisisalmah52@gmail.com, Nigeria, Data Glacier, Data Analyst

Buse Gungor, busegungor2303@gmail.com, Turkey, Data Glacier, Data Analyst

Problem Description

XYZ credit union in Latin America is performing very well in selling the Banking products (eg: Credit card, deposit account, retirement account, safe deposit box etc.) but their existing customer is not buying more than 1 product which means bank is not performing good in cross selling (Bank is not able to sell their other offerings to existing customer). As ABC analytics firm, need to inspect the data and suggest what action bank can take to increase cross selling without using ML.

Data Understanding

We are provided with 1.5 years of customers behaviour data from a bank to predict what new products customers will purchase. The data starts at 2015-01-28 and has monthly records of products a customer has, such as "credit card", "savings account", etc. We will predict what additional products a customer will get in the last month, 2016-06-28, in addition to what they already have at 2016-05-28. These products are the columns named: ind_(xyz)_ult1, which are the columns #25 - #48 in the training data. You will predict what a customer will buy in addition to what they already had at 2016-05-28.
The test and train sets are split by time, and public and private leader board sets are split randomly.

File descriptions

- train.csv - the training set
- test.csv - the test set

What are the problems in data?

Two columns have many Null values that will cause "Nans-Imputation" problems, so we deleted them. We also deleted "sexo", "canal_entrada", "segmento" and "nomprov" columns for the same reason.
Based on our EDA, some columns have no relevance to the model that we planned and we deleted them as well, such as:
"cod_prov",'ind_actividad_cliente',"indrel_1mes",'indresi',"tipodom","ind_empleado" ,"pais_residencia","indrel","indext","indfall","ind_nuevo"

Furthermore, we deleted products that nobody buys:
'ind_ahor_fin_ult1','ind_aval_fin_ult1','ind_cder_fin_ult1','ind_ctju_fin_ult1','ind_deco_fin_ult1','ind_deme_fin_ult1', 'ind_pres_fin_ult1','ind_viv_fin_ult1'

We decided to follow an alternative approach to deal with missing values; we used the most frequent imputation strategy for the replacement, then we reapplied the previous filter to prevent unexpected values.

Lastly, we changed the data types of numeric columns to "int32" and filtered to get clients without age outliers values between 18 and 100 years old.

We binned the “canal entrada” by keeping the three biggest values and converting the rest to the same unique group “Others.”