

```
In [1]: #Importing Libraries
import os
os.chdir('C:/Users/TOTAG005884/Documents/Totago Technologies/Data Science/Projects/BOG')

In [2]: #Importing Libraries
import pandas as pd
import numpy as np
from sklearn import preprocessing
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn import preprocessing
from sklearn.datasets import make_classification
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import RepeatedStratifiedKFold
from sklearn.linear_model import LogisticRegression
from sklearn.neighbours import KNeighborsClassifier
from sklearn.metrics import decision_tree_classification
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.ensemble import StackingClassifier
from matplotlib import pyplot
from numpy import mean
from numpy import std
from sklearn.ensemble import AdaBoostClassifier
from xgboost import XGBClassifier
from lightgbm import LGBMClassifier
from catboost import CatBoostClassifier
from sklearn.metrics import roc_auc_score, log_loss
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import precision_recall_fscore_support as score
from sklearn.model_selection import train_test_split
from sklearn.ensemble import GradientBoostingClassifier
```

Datasets

```
In [3]: import pandas as pd
cust_data = pd.read_csv('ml_case_training_data.csv')
price_data = pd.read_csv('ml_case_training_hist_data.csv')
forecast_cons = pd.read_csv('ml_case_training_output.csv')
```

```
In [4]: #Customer Dataset
cust_data.head(10)
```

```
Out[4]:
```

	id	activity_new	campaign_disc_ele	channel_sales	cons_12m
0	48a8a52261e7c58715202705a0451c9	essolipkbfksuakmfwucbzccomxiw	NaN	lmkebamcaadubkfadmueccozcimema	309275
1	24011ae4eb303511d65fa7c15bc57	foosdfyfuasacimwkcsozibcdkciaua	NaN	foosdfyfuasacimwkcsozibcdkciaua	0
2	d29c2c54ac38f3c06144da0a53813dd	NaN	NaN	NaN	4680
3	764c75961154da3ac3a6c254cd082ea7d	NaN	NaN	foosdfyfuasacimwkcsozibcdkciaua	544
4	bb4a3439a252a1e16080264c16191cb	NaN	NaN	lmkebamcaadubkfadmueccozcimema	1584
5	568bb38a1af7c0f4dc77b378959a3	sfsdfscdfpmckuekkuksaeixdaeu	NaN	foosdfyfuasacimwkcsozibcdkciaua	121335
6	149d57df26c1d494415803a877cb4b	NaN	NaN	NaN	4425
7	1aa08825382410a098937d5504ec26d	NaN	NaN	usluppasemubtkpaafesmlbmzsf	8302
8	7ab4b4878d8f7661fd2c0b98a18011	sscdlpkpfkpfkuekkuksaeixdaeu	NaN	foosdfyfuasacimwkcsozibcdkciaua	45097
9	01495c955ba7ec5e7673203406755aa0	NaN	NaN	foosdfyfuasacimwkcsozibcdkciaua	29552

10 rows × 5 columns

```
In [5]: price_data.head(10)
```

```
Out[5]:
```

	id	price_date	price_p1_var	price_p2_var	price_p3_var	price_p1_fix	price_p2_fix	price_p3_fix
0	038af19179925da21a25619c5a24b745	2015-01-01	0.151367	0.0	0.0	44.266931	0.0	0.0
1	038af19179925da21a25619c5a24b745	2015-03-01	0.151367	0.0	0.0	44.266931	0.0	0.0
2	038af19179925da21a25619c5a24b745	2015-03-01	0.151367	0.0	0.0	44.266931	0.0	0.0
4	038af19179925da21a25619c5a24b745	2015-05-01	0.149626	0.0	0.0	44.266931	0.0	0.0
5	038af19179925da21a25619c5a24b745	2015-05-01	0.149626	0.0	0.0	44.266931	0.0	0.0
6	038af19179925da21a25619c5a24b745	2015-07-01	0.150321	0.0	0.0	44.444710	0.0	0.0
7	038af19179925da21a25619c5a24b745	2015-08-01	0.148599	0.0	0.0	44.444710	0.0	0.0
8	038af19179925da21a25619c5a24b745	2015-09-01	0.148599	0.0	0.0	44.444710	0.0	0.0
9	038af19179925da21a25619c5a24b745	2015-10-01	0.148599	0.0	0.0	44.444710	0.0	0.0

```
In [6]: #Churn Data
churn_data.head(10)
```

```
Out[6]:
```

	id	churn
0	48a8a52261e7c58715202705a0451c9	0
1	24011ae4eb303511d65fa7c15bc57	1
2	d29c2c54ac38f3c06144da0a53813dd	0
3	764c75961154da3ac3a6c254cd082ea7d	0
4	bb4a3439a252a1e16080264c16191cb	0
5	568bb38a1af7c0f4dc77b378959a3	0
6	149d57df26c1d494415803a877cb4b	0
7	1aa08825382410a098937d5504ec26d	1
8	7ab4b4878d8f7661fd2c0b98a18011	1
9	01495c955ba7ec5e7673203406755aa0	0

Data Cleaning

Dealing with missing data

```
In [7]: #Checking for null values in Customer Details Dataset
cust_data.isnull().sum()
```

```
Out[7]:
```

id	0
activity_new	9545
campaign_disc_ele	16096
channel_sales	4218
cons_12m	0
cons_gas_12m	0
cons_last_month	0
date_activ	0
date_end	2
date_first_activ	12588
date_modif_prod	157
date_renewal	40
forecast_base_bill_ele	12588
forecast_base_bill_year	12588
forecast_bill_12m	12588
forecast_cons	12588
forecast_cons_12m	0
forecast_cons_year	0
forecast_discount_energy	126
forecast_meter_rent_12m	0
forecast_price_energy_p1	126
forecast_price_energy_p2	126
forecast_price_pow_p1	126
has_gas	0
imp_cons	13
margin_gross_pow_ele	0
margin_net_pow_ele	13
nb_prod_act	0
net_margin	15
num_years_antig	0
origin_up	87
pow_max	3
dtype:	int64

Columns with null values = activity_new, campaign_disc_ele, channel_sales, date_end, date_first_activ, date_modif_prod, date_renewal, forecast_base_bill_ele, forecast_base_bill_year, forecast_bill_12m, forecast_cons, forecast_discount_energy, forecast_price_energy_p1, forecast_price_energy_p2, forecast_price_pow_p1, margin_gross_pow_ele, margin_net_pow_ele, net_margin, origin_up, pow_max

```
In [8]: #Checking for null values in Price Dataset
price_data.isnull().sum()
```

```
Out[8]:
```

id	0
price_date	0
price_p1_var	1359
price_p2_var	1359
price_p3_var	1359
price_p1_fix	1359
price_p2_fix	1359
price_p3_fix	1359
dtype:	int64

Columns with missing data = price_p1_var, price_p2_var, price_p3_var, price_p1_fix, price_p2_fix, price_p3_fix

```
In [9]: #Checking for null values in Churn Dataset
churn_data.isnull().sum()
```

```
Out[9]:
```

id	0
churn	0
dtype:	int64

```
In [10]: #Converting the dates to datetime
import datetime
for a in ['date_activ', 'date_end', 'date_first_activ', 'date_modif_prod', 'date_renewal']:
    cust_data[a] = pd.to_datetime(cust_data[a], format='%Y %m %d')
cust_data.head(10)
```

```
Out[10]:
```

	id	activity_new	campaign_disc_ele	channel_sales	cons_12m
0	48a8a52261e7c58715202705a0451c9	essolipkbfksuakmfwucbzccomxiw	NaN	lmkebamcaadubkfadmueccozcimema	309275
1	24011ae4eb303511d65fa7c15bc57	NaN	NaN	foosdfyfuasacimwkcsozibcdkciaua	0
2	d29c2c54ac38f3c06144da0a53813dd	NaN	NaN	NaN	4680
3	764c75961154da3ac3a6c254cd082ea7d	NaN	NaN	foosdfyfuasacimwkcsozibcdkciaua	544
4	bb4a3439a252a1e16080264c16191cb	NaN	NaN	lmkebamcaadubkfadmueccozcimema	1584
5	568bb38a1af7c0f4dc77b378959a3	sfsdfscdfpmckuekkuksaeixdaeu	NaN	foosdfyfuasacimwkcsozibcdkciaua	121335
6	149d57df26c1d494415803a877cb4b	NaN	NaN	NaN	4425
7	1aa08825382410a098937d5504ec26d	NaN	NaN	usluppasemubtkpaafesmlbmzsf	8302
8	7ab4b4878d8f7661fd2c0b98a18011	sscdlpkpfkpfkuekkuksaeixdaeu	NaN	foosdfyfuasacimwkcsozibcdkciaua	45097
9	01495c955ba7ec5e7673203406755aa0	NaN	NaN	foosdfyfuasacimwkcsozibcdkciaua	29552

10 rows × 5 columns

No nulls here

```
In [11]: #Filling up nulls in Customer Details Dataset
#Filling with zeros
for col in ['forecast_base_bill_ele', 'forecast_base_bill_year', 'forecast_bill_12m', 'forecast_cons',
'forecast_discount_energy', 'forecast_price_energy_p1', 'forecast_price_energy_p2', 'forecast_price_pow_p1',
'forecast_price_energy_p2', 'margin_gross_pow_ele', 'margin_net_pow_ele', 'net_margin', 'pow_max']:
    cust_data[col].fillna(0, inplace = True)

#Filling with null keyword
for col in ['channel_sales', 'origin_up']:
    cust_data[col].fillna('Null', inplace = True)

#Filling dates with activation date
for col in ['date_activ', 'date_first_activ', 'date_modif_prod', 'date_renewal']:
    cust_data[col].fillna(cust_data['date_activ'], inplace = True)

#Dropping Some columns
cust_data.drop(columns=['activity_new', 'campaign_disc_ele', ], inplace=True)
cust_data.head(10)
```

```
Out[11]:
```

	id	channel_sales	cons_12m	cons_gas_12m	cons_last_month	date_activ	date_e
0	48a8a52261e7c58715202705a0451c9	lmkebamcaadubkfadmueccozcimema	309275	0	10025	2012-11-16	2016-01-07
1	24011ae4eb303511d65fa7c15bc57	foosdfyfuasacimwkcsozibcdkciaua	0	54946	0	2013-06-15	2016-01-08
2	d29c2c54ac38f3c06144da0a53813dd	Null	4680	0	0	2009-08-21	2016-01-15
3	764c75961154da3ac3a6c254cd082ea7d	foosdfyfuasacimwkcsozibcdkciaua	544	0	0	2010-04-16	2016-01-16
4	bb4a3439a252a1e16080264c16191cb	lmkebamcaadubkfadmueccozcimema	1584	0	0	2010-03-30	2016-01-16
5	568bb38a1af7c0f4dc77b378959a3	foosdfyfuasacimwkcsozibcdkciaua	121335	0	12400	2010-04-08	2016-01-16
6	149d57df26c1d494415803a877cb4b	Null	4425	0	526	2010-01-13	2016-01-16
7	1aa08825382410a098937d5504ec26d	usluppasemubtkpaafesmlbmzsf	8302	0	1998	2011-12-09	2016-01-16
8	7ab4b4878d8f7661fd2c0b98a18011	foosdfyfuasacimwkcsozibcdkciaua	45097	0	0	2011-12-02	2016-01-16
9	01495c955ba7ec5e7673203406755aa0	foosdfyfuasacimwkcsozibcdkciaua	29552	0	1260	2010-04-21	2016-01-16

10 rows × 7 columns

```
In [12]: #Filling up nulls in Price Dataset
for col in ['price_p1_var', 'price_p2_var', 'price_p3_var', 'price_p1_fix', 'price_p2_fix', 'price_p3_fix']:
    price_data[col].fillna(0, inplace = True)
price_data.head(10)
```

```
Out[12]:
```

	id	price_p1_var	price_p2_var	price_p3_var	price_p1_fix	price_p2_fix	price_p3_fix	
0	038af19179925da21a25619c5a24b745	2015-01-01	0.151367	0.0	0.0	44.266931	0.0	0.0
1	038af19179925da21a25619c5a24b745	2015-02-01	0.151367	0.0	0.0	44.266931	0.0	0.0
2	038af19179925da21a25619c5a24b745	2015-03-01	0.151367	0.0	0.0	44.266931	0.0	0.0
3	038af19179925da21a25619c5a24b745	2015-04-01	0.149626	0.0	0.0	44.266931	0.0	0.0
4	038af19179925da21a25619c5a24b745	2015-05-01	0.149626	0.0	0.0	44.266931	0.0	0.0
5	038af19179925da21a25619c5a24b745	2015-06-01	0.149626	0.0	0.0	44.266930	0.0	0.0
6	038af19179925da21a25619c5a24b745	2015-07-01	0.150321	0.0	0.0	44.444710	0.0	0.0
7	038af19179925da21a25619c5a24b745	2015-08-01	0.148599	0.0	0.0	44.444710	0.0	0.0
8	038af19179925da21a25619c5a24b745	2015-09-01	0.148599	0.0	0.0	44.444710	0.0	0.0
9	038af19179925da21a25619c5a24b745	2015-10-01	0.148599	0.0	0.0	44.444710	0.0	0.0

```
In [13]: #Cross checking nulls and checking variable types for Customer Details Dataset
cust_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16096 entries, 0 to 16095
Data columns (total 30 columns):
# Column Non-Null Count Dtype
---  ---
0 id 16096 non-null object
1 channel_sales 16096 non-null object
2 cons_12m 16096 non-null int64
3 cons_gas_12m 16096 non-null int64
4 cons_last_month 16096 non-null object
5 date_activ 16096 non-null datetime64[ns]
6 date_end 16096 non-null datetime64[ns]
7 date_first_activ 16096 non-null datetime64[ns]
8 date_modif_prod 16096 non-null datetime64[ns]
9 date_renewal 16096 non-null datetime64[ns]
10 forecast_base_bill_ele 16096 non-null float64
11 forecast_base_bill_year 16096 non-null float64
12 forecast_bill_12m 16096 non-null float64
13 forecast_cons 16096 non-null float64
14 forecast_cons_12m 16096 non-null float64
15 forecast_cons_year 16096 non-null int64
16 forecast_discount_energy 16096 non-null float64
17 forecast_meter_rent_12m 16096 non-null float64
18 forecast_price_energy_p1 16096 non-null float64
19 forecast_price_energy_p2 16096 non-null float64
20 forecast_price_pow_p1 16096 non-null float64
21 has_gas 16096 non-null object
22 imp_cons 16096 non-null float64
23 margin_gross_pow_ele 16096 non-null float64
24 margin_net_pow_ele 16096 non-null float64
25 nb_prod_act 16096 non-null int64
26 net_margin 16096 non-null float64
27 num_years_antig 16096 non-null int64
28 origin_up 16096 non-null object
29 pow_max 16096 non-null float64
dtypes: datetime64[ns](15), float64(15), int64(6), object(4)
memory usage: 3.7+ MB
```

```
In [14]: #Cross checking nulls and checking variable types for Price Dataset
price_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 193002 entries, 0 to 193001
Data columns (total 8 columns):
# Column Non-Null Count Dtype
---  ---
0 id 193002 non-null object
1 price_date 193002 non-null object
2 price_p1_var 193002 non-null float64
3 price_p2_var 193002 non-null float64
4 price_p3_var 193002 non-null float64
5 price_p1_fix 193002 non-null float64
6 price_p2_fix 193002 non-null float64
7 price_p3_fix 193002 non-null float64
dtypes: float64(6), object(2)
memory usage: 11.8+ MB
```

Merge Data

```
In [20]: # Inner join customer data and churn data
Data = pd.merge(cust_data, churn_data, left_on = 'id', right_on = 'id')
Data.head(10)
```

```
Out[20]:
```

	id	channel_sales	cons_12m	cons_gas_12m	cons_last_month	date_activ	date_e
0	48a8a52261e7c58715202705a0451c9	lmkebamcaadubkfadmueccozcimema	309275	0	10025	2012-11-16	2016-01-07
1	24011ae4eb303511d65fa7c15bc57	foosdfyfuasacimwkcsozibcdkciaua	0	54946	0	2013-06-15	2016-01-08
2	d29c2c54ac38f3c06144da0a53813dd	Null	4680	0	0	2009-08-21	2016-01-15
3	764c75961154da3ac3a6c254cd082ea7d	foosdfyfuasacimwkcsozibcdkciaua	544	0	0	2010-04-16	2016-01-16
4	bb4a3439a252a1e16080264c16191cb	lmkebamcaadubkfadmueccozcimema	1584	0	0	2010-03-30	2016-01-16
5	568bb38a1af7c0f4dc77b378959a3	foosdfyfuasacimwkcsozibcdkciaua	121335	0	12400	2010-04-08	2016-01-16
6	149d57df26c1d494415803a877cb4b	Null	4425	0	526	2010-01-13	2016-01-16
7	1aa08825382410a098937d5504ec26d	usluppasemubtkpaafesmlbmzsf	8302	0	1998	2011-12-09	2016-01-16
8	7ab4b4878d8f7661fd2c0b98a18011	foosdfyfuasacimwkcsozibcdkciaua	45097	0	0	2011-12-02	2016-01-16
9	01495c955ba7ec5e7673203406755aa0	foosdfyfuasacimwkcsozibcdkciaua	29552	0	1260	2010-04-21	2016-01-16

10 rows × 7 columns

```
In [21]: #dropping the price date, so as to group the data by id
price_dataset = price_data.drop(['price_date'], axis = 1)

#grouping the data by id, and summing the prices
new_price_data = price_dataset.groupby(['id']).sum()
new_price_data.head(10)
```

```
Out[21]:
```

	id	price_p1_var	price_p2_var	price_p3_var	price_p1_fix	price_p2_fix	price_p3_fix
0004230fbb638288a6329a6628c38d	1.420061	1.245525	0.877924	488.420778	283.052462	195.368327	
0004230fbb638288a6329a6628c38d	1.757118	0.000000	0.000000	532.625404	0.000000	0.000000	
0010bc39e423c21f1ed2c855a2463e	2.178702	0.000000	0.000000	543.836520	0.000000	0.000000	
001f463855f9a67477d79b8a79a08537	1.425085	1.179599	0.828384	487.769130	292.661462	195.107656	
00126c47d7967679b79b9a79a08537	1.775110	0.000000	0.000000	531.203164	0.000000	0.000000	
001326a539a26c79a679b1869934227	1.437078	1.193008	0.843051	487.932042	282.792414	195.172828	
0019325a539a26c79a679b1869934227	1.512911	1.266503	0.899048	488.746620	293.247962	195.498660	
001f5e22779e7f33a75636a084006	1.771648	0.000000	0.000000	531.203164	0.000000	0.000000	
001987d9b0ba4e4a7238a77233b1f4	1.473067	1.227485	0.876388	487.769128	292.661464	195.107662	
0019af3c2142c2d9a69203a446f	3.209391	0.000000	0.000000	695.543164	0.000000	0.000000	

```
In [22]: #Join id to Price dates
Dataset = pd.merge(Data, new_price_data, left_on = 'id', right_on = 'id', indicator = True)
Dataset.head(10)
```

```
Out[22]:
```

	id	channel_sales	cons_12m	cons_gas_12m</
--	----	---------------	----------	----------------

