In [1]: ##Setting directory import os os.chdir('C:/Users/TOTAGOUSER4/Documents/Totago Technologies/Data Science/Projects/BCG') In [2]: #Importing Libraries import pandas as pd import numpy as np import time 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.neighbors import KNeighborsClassifier from sklearn.tree import DecisionTreeClassifier 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** import pandas as pd In [3]: cust data = pd.read csv('ml case training data.csv') price_data = pd.read_csv('ml_case_training_hist_data.csv') churn_data = pd.read_csv('ml_case_training_output.csv') #Customer Dataset cust_data.head(10) Out[4]: id activity_new campaign_disc_ele channel_sales cons_12m 48ada52261e7cf58715202705a0451c9 esoiiifxdlbkcsluxmfuacbdckommixw Imkebamcaaclubfxadlmueccxoimlema 309275 24011ae4ebbe3035111d65fa7c15bc57 NaN NaN foosdfpfkusacimwkcsosbicdxkicaua d29c2c54acc38ff3c0614d0a653813dd NaN NaN NaN 4660 764c75f661154dac3a6c254cd082ea7d NaN NaN foosdfpfkusacimwkcsosbicdxkicaua 544 Imkebamcaaclubfxadlmueccxoimlema bba03439a292a1e166f80264c16191cb NaN 1584 568bb38a1afd7c0fc49c77b3789b59a3 sfisfxfcocfpcmckuekokxuseixdaoeu NaN foosdfpfkusacimwkcsosbicdxkicaua 121335 149d57cf92fc41cf94415803a877cb4b NaN NaN NaN 4425 7 1aa498825382410b098937d65c4ec26d NaN NaN usilxuppasemubllopkaafesmlibmsdf 8302 7ab4bf4878d8f7661dfc20e9b8e18011 sscfoipxikopfskekuobeuxkxmwsuucb foosdfpfkusacimwkcsosbicdxkicaua NaN 45097 NaN NaN 01495c955be7ec5e7f3203406785aae0 foosdfpfkusacimwkcsosbicdxkicaua 29552 10 rows × 32 columns In [5]: #Pricing Data price_data.head(10) Out[5]: 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 0.0 0.0 44.266931 0.0 0.0 0.151367 1 038af19179925da21a25619c5a24b745 2015-02-01 0.151367 0.0 0.0 44.266931 0.0 0.0 2 038af19179925da21a25619c5a24b745 2015-03-01 0.0 0.0 44.266931 0.0 0.0 0.151367 3 038af19179925da21a25619c5a24b745 2015-04-01 0.149626 0.0 0.0 44.266931 0.0 0.0 0.0 0.0 4 038af19179925da21a25619c5a24b745 2015-05-01 0.149626 0.0 44.266931 0.0 038af19179925da21a25619c5a24b745 2015-06-01 0.149626 0.0 0.0 44.266930 0.0 0.0 0.0 6 038af19179925da21a25619c5a24b745 2015-07-01 0.150321 0.0 44.444710 0.0 0.0 **7** 038af19179925da21a25619c5a24b745 2015-08-01 0.145859 0.0 0.0 44.444710 0.0 0.0 0.0 0.0 0.0 038af19179925da21a25619c5a24b745 2015-09-01 0.145859 0.0 44.444710 9 038af19179925da21a25619c5a24b745 2015-10-01 0.145859 0.0 0.0 44.444710 0.0 0.0 In [6]: #Churn data churn_data.head(10) Out[6]: id churn 48ada52261e7cf58715202705a0451c9 0 24011ae4ebbe3035111d65fa7c15bc57 2 d29c2c54acc38ff3c0614d0a653813dd 764c75f661154dac3a6c254cd082ea7d 3 bba03439a292a1e166f80264c16191cb 568bb38a1afd7c0fc49c77b3789b59a3 6 149d57cf92fc41cf94415803a877cb4b 1aa498825382410b098937d65c4ec26d 7ab4bf4878d8f7661dfc20e9b8e18011 01495c955be7ec5e7f3203406785aae0 **Data Cleaning** Dealing with missing data #Checking for null values in Customer Details Dataset In [7]: cust data.isnull().sum() Out[7]: id 0 activity new 9545 campaign disc ele 16096 channel sales 4218 cons 12m cons gas 12m 0 0 cons last month date activ 2 date end date_first_activ 12588 157 date modif prod 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 126 forecast discount energy forecast meter rent 12m 0 126 forecast_price_energy_p1 126 forecast_price_energy_p2 126 forecast_price_pow_p1 has gas 0 imp cons 0 13 margin_gross_pow_ele 13 margin net pow ele nb prod act 0 15 net margin num_years_antig 0 87 origin up 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 1359 price_p1_var price p2 var 1359 price_p3_var 1359 price_p1_fix 1359 1359 price_p2_fix 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 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]: activity_new campaign_disc_ele channel_sales cons_12m 48ada52261e7cf58715202705a0451c9 esoiiifxdlbkcsluxmfuacbdckommixw NaN Imkebamcaaclubfxadlmueccxoimlema 309275 24011ae4ebbe3035111d65fa7c15bc57 NaN NaN foosdfpfkusacimwkcsosbicdxkicaua 0 d29c2c54acc38ff3c0614d0a653813dd 2 NaN NaN NaN 4660 764c75f661154dac3a6c254cd082ea7d NaN NaN foosdfpfkusacimwkcsosbicdxkicaua 544 bba03439a292a1e166f80264c16191cb NaN NaN Imkebamcaaclubfxadlmueccxoimlema 1584 5 568bb38a1afd7c0fc49c77b3789b59a3 sfisfxfcocfpcmckuekokxuseixdaoeu NaN foosdfpfkusacimwkcsosbicdxkicaua 121335 149d57cf92fc41cf94415803a877cb4b NaN NaN NaN 4425 1aa498825382410b098937d65c4ec26d NaN NaN usilxuppasemubllopkaafesmlibmsdf 8302 7ab4bf4878d8f7661dfc20e9b8e18011 sscfoipxikopfskekuobeuxkxmwsuucb NaN foosdfpfkusacimwkcsosbicdxkicaua 45097 01495c955be7ec5e7f3203406785aae0 NaN NaN foosdfpfkusacimwkcsosbicdxkicaua 29552 10 rows × 32 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', '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_end', '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_eı 2016-1 2012-11-48ada52261e7cf58715202705a0451c9 ImkebamcaaclubfxadImueccxoimlema 309275 10025 07 2013-06-2016-0 24011ae4ebbe3035111d65fa7c15bc57 foosdfpfkusacimwkcsosbicdxkicaua 54946 15 2009-08-2016-0 d29c2c54acc38ff3c0614d0a653813dd Null 4660 2010-04-2016-0 764c75f661154dac3a6c254cd082ea7d foosdfpfkusacimwkcsosbicdxkicaua 544 0 16 2010-03-2016-0 bba03439a292a1e166f80264c16191cb Imkebamcaaclubfxadlmueccxoimlema 1584 30 2010-04-2016-0 5 568bb38a1afd7c0fc49c77b3789b59a3 foosdfpfkusacimwkcsosbicdxkicaua 121335 12400 08 2016-0 2010-01-149d57cf92fc41cf94415803a877cb4b Null 4425 526 13 2011-12-2016-1 1aa498825382410b098937d65c4ec26d usilxuppasemubllopkaafesmlibmsdf 8302 0 1998 2016-1 2011-12-7ab4bf4878d8f7661dfc20e9b8e18011 foosdfpfkusacimwkcsosbicdxkicaua 45097 02 2010-04-2016-0 01495c955be7ec5e7f3203406785aae0 foosdfpfkusacimwkcsosbicdxkicaua 29552 1260 21 10 rows × 30 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_f ix']: price_data[col].fillna(0, inplace = True) price data.head(10) Out[12]: 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 0.151367 **1** 038af19179925da21a25619c5a24b745 2015-02-01 0.0 0.0 44.266931 0.0 0.0 2 038af19179925da21a25619c5a24b745 2015-03-01 44.266931 0.151367 0.0 0.0 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 0.149626 **5** 038af19179925da21a25619c5a24b745 2015-06-01 0.0 0.0 44.266930 0.0 0.0 44.444710 6 038af19179925da21a25619c5a24b745 2015-07-01 0.0 0.0 0.150321 0.0 0.0 **7** 038af19179925da21a25619c5a24b745 2015-08-01 0.145859 0.0 0.0 44.444710 0.0 0.0 8 038af19179925da21a25619c5a24b745 2015-09-01 44.444710 0.145859 0.0 0.0 0.0 0.0 9 038af19179925da21a25619c5a24b745 2015-10-01 0.145859 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 16096 non-null object 1 channel sales 16096 non-null int64 2 cons 12m cons gas 12m 16096 non-null int64 3 16096 non-null int64 4 cons_last_month 5 date activ 16096 non-null datetime64[ns] 16096 non-null datetime64[ns]
16096 non-null datetime64[ns]
16096 non-null datetime64[ns] 6 date end 7 date first activ date modif prod 16096 non-null datetime64[ns] date renewal 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 15 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 16096 non-null float64 26 net_margin 27 num_years_antig 16096 non-null int64 28 origin_up 16096 non-null object 16096 non-null float64 29 pow max dtypes: datetime64[ns](5), 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 price date 193002 non-null object 2 price p1 var 193002 non-null float64 price p2 var 193002 non-null float64 price p3 var 193002 non-null 5 price_p1_fix 193002 non-null 6 price_p2_fix 193002 non-null float64 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]: channel_sales cons_12m cons_gas_12m cons_last_month date_activ date_er 2012-11-48ada52261e7cf58715202705a0451c9 ImkebamcaaclubfxadImueccxoimlema 309275 0 10025 07 2013-06-2016-0 24011ae4ebbe3035111d65fa7c15bc57 54946 foosdfpfkusacimwkcsosbicdxkicaua 15 2009-08-2016-0 d29c2c54acc38ff3c0614d0a653813dd Null 4660 21 2010-04-2016-0 0 764c75f661154dac3a6c254cd082ea7d foosdfpfkusacimwkcsosbicdxkicaua 544 16 2010-03-2016-0 bba03439a292a1e166f80264c16191cb Imkebamcaaclubfxadlmueccxoimlema 1584 30 2010-04-568bb38a1afd7c0fc49c77b3789b59a3 foosdfpfkusacimwkcsosbicdxkicaua 121335 0 12400 80 2016-0 2010-01-149d57cf92fc41cf94415803a877cb4b 4425 526 13 2011-12-1aa498825382410b098937d65c4ec26d usilxuppasemubllopkaafesmlibmsdf 8302 0 1998 09 2016-1 2011-12-7ab4bf4878d8f7661dfc20e9b8e18011 foosdfpfkusacimwkcsosbicdxkicaua 45097 02 2010-04-01495c955be7ec5e7f3203406785aae0 foosdfpfkusacimwkcsosbicdxkicaua 29552 0 1260 21 10 rows × 31 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]: price_p1_var price_p2_var price_p3_var price_p1_fix price_p2_fix price_p3_fix id 0002203ffbb812588b632b9e628cc38d 1.492061 1.245525 0.877924 488.420789 293.052462 195.368327 0004351ebdd665e6ee664792efc4fd13 1.757118 0.000000 0.000000 532.625404 0.000000 0.000000 0010bcc39e42b3c2131ed2ce55246e3c 2.178702 0.000000 0.000000 543.836520 0.000000 0.000000 0010ee3855fdea87602a5b7aba8e42de 1.425085 1.179509 0.828384 487.769130 292.661462 195.107656 00114d74e963e47177db89bc70108537 1.775110 0.000000 0.000000 531.203164 0.000000 0.000000 00126c87cf78d7604278f0a9adeb689e 1.437678 0.843651 487.932042 292.759214 1.193008 195.172828 0013f326a839a2f6ad87a1859952d227 1.512911 1.266503 0.899048 488.746620 293.247960 195.498660 00184e957277eeef733a7b563fdabd06 1.771648 0.000000 0.000000 531.203164 0.000000 0.000000 001987ed9dbdab4efa274a9c7233e1f4 1.473067 1.227485 0.876366 487.769126 292.661464 195.107662 0019baf3ed1242cd99b3cb592030446f 3.209391 0.000000 0.000000 0.000000 0.000000 695.543164 In [22]: #Join it to Price data. Dataset = pd.merge(Data, new_price_data, left_on = 'id', right_on = 'id', indicator = True) Dataset.head(10) Out[22]: date_eı id channel_sales cons_12m cons_gas_12m cons_last_month date_activ 2012-11-2016-1 48ada52261e7cf58715202705a0451c9 Imkebamcaaclubfxadlmueccxoimlema 309275 0 10025 07 2013-06-2016-0 0 24011ae4ebbe3035111d65fa7c15bc57 foosdfpfkusacimwkcsosbicdxkicaua 0 54946 15 2009-08-2016-0 d29c2c54acc38ff3c0614d0a653813dd Null 4660 0 0 21 2010-04-2016-0 764c75f661154dac3a6c254cd082ea7d foosdfpfkusacimwkcsosbicdxkicaua 544 0 0 16 2010-03-2016-0 bba03439a292a1e166f80264c16191cb Imkebamcaaclubfxadlmueccxoimlema 0 1584 30 2010-04-2016-0 0 12400 5 568bb38a1afd7c0fc49c77b3789b59a3 foosdfpfkusacimwkcsosbicdxkicaua 121335 80 2010-01-2016-0 Null 4425 526 149d57cf92fc41cf94415803a877cb4b 13 2011-12-2016-1 1aa498825382410b098937d65c4ec26d usilxuppasemubllopkaafesmlibmsdf 8302 0 1998 09 2011-12-2016-1 7ab4bf4878d8f7661dfc20e9b8e18011 foosdfpfkusacimwkcsosbicdxkicaua 45097 02 2010-04-2016-0 0 01495c955be7ec5e7f3203406785aae0 foosdfpfkusacimwkcsosbicdxkicaua 29552 1260 21 10 rows × 38 columns **Feature Engineering** Dataset.isnull().sum() In [23]: Out[23]: id 0 0 channel_sales cons 12m 0 cons gas 12m 0 cons last_month 0 0 date activ 0 date end date first activ 0 date modif prod 0 0 date renewal 0 forecast base bill ele forecast base bill year 0 forecast bill 12m 0 0 forecast cons forecast cons 12m forecast_cons_year forecast discount energy forecast meter rent 12m forecast price energy p1 forecast price energy p2 forecast_price_pow_p1 0 has gas imp cons margin gross pow ele margin_net_pow_ele nb prod act 0 net margin num years antig 0 origin_up 0 pow max 0 churn price_p1_var price_p2_var price p3 var 0 price p1 fix 0 0 price_p2_fix 0 price_p3_fix merge dtype: int64 no null values In [24]: Dataset.info() <class 'pandas.core.frame.DataFrame'> Int64Index: 16096 entries, 0 to 16095 Data columns (total 38 columns): # Column Non-Null Count Dtype 0 id 16096 non-null object 16096 non-null object 1 channel sales cons 12m 16096 non-null int64 3 cons_gas_12m 16096 non-null int64 4 16096 non-null int64 cons last month 5 date activ 16096 non-null datetime64[ns] 16096 non-null datetime64[ns] 6 date end date_first_activ 16096 non-null datetime64[ns] 8 16096 non-null datetime64[ns] date modif prod 16096 non-null datetime64[ns] 9 date renewal 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 16096 non-null int64 15 forecast_cons_year 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 16096 non-null float64 24 margin_net_pow_ele 25 nb_prod_act 16096 non-null int64 16096 non-null float64 26 net_margin 16096 non-null int64 27 num_years_antig 16096 non-null object 28 origin_up 29 pow_max 16096 non-null float64 30 churn 16096 non-null int64 31 price_p1_var 16096 non-null float64 32 price_p2_var 16096 non-null float64 16096 non-null float64 33 price p3 var 34 16096 non-null price_p1_fix 16096 non-null float64 35 price_p2_fix 36 price_p3_fix 16096 non-null float64 16096 non-null category 37 merge dtypes: category(1), datetime64[ns](5), float64(21), int64(7), object(4) memory usage: 4.7+ MB Data types are fine Now, Let's generate some features. Feature needed; Those present are; channel_sales, cons_12m, cons_gas_12m, cons_gas_12m, cons_gas_12m, cons_last_month, forecast_cons_12m, forecast_cons_year, forecast_discount_energy, forecast_meter_rent_12m, forecast_price_energy_p1, forecast_price_energy_p2, forecast_price_pow_p1, has_gas, imp_cons, margin_gross_pow_ele, margin_net_pow_ele, nb_prod_act, net_margin, num_years_antig, origin_up, pow_max, price_p1_var, price_p2_var, price_p3_var, price_p1_fix, price_p2_fix, price_p3_fix Those to be generated; num_months_antig (number of months the Customer has spent), num_years_renewal (time before contract renewal), avg_month_elec (Average consuption of electricity per month), avg_month_gas (Average consumption of gas per month), #generating Average consuption of electricity per month In [25]: Dataset['avg_month_elec'] = round(Dataset['cons_12m']/12, 2) #generating Average consumption of gas per month Dataset['avg_month_gas'] = round(Dataset['cons_gas_12m']/12, 2) # generating time before contract renewal Dataset['num_years_renewal'] = ((Dataset['date_renewal']).dt.year - (Dataset['date_activ']).dt.year) Dataset.head(10) Out[25]: channel_sales cons_12m cons_gas_12m cons_last_month date_activ date_eı 2016-1 2012-11-48ada52261e7cf58715202705a0451c9 Imkebamcaaclubfxadlmueccxoimlema 10025 309275 07 2013-06-2016-0 24011ae4ebbe3035111d65fa7c15bc57 foosdfpfkusacimwkcsosbicdxkicaua 54946 15 2016-0 2009-08d29c2c54acc38ff3c0614d0a653813dd Null 4660 0 21 2010-04-2016-0 764c75f661154dac3a6c254cd082ea7d foosdfpfkusacimwkcsosbicdxkicaua 544 0 16 2010-03-2016-0 bba03439a292a1e166f80264c16191cb Imkebamcaaclubfxadlmueccxoimlema 1584 0 30 2010-04-2016-0 568bb38a1afd7c0fc49c77b3789b59a3 foosdfpfkusacimwkcsosbicdxkicaua 121335 12400 08 2010-01-2016-0 149d57cf92fc41cf94415803a877cb4b 6 Null 4425 0 526 13 2011-12-2016-1 1aa498825382410b098937d65c4ec26d usilxuppasemubllopkaafesmlibmsdf 8302 0 1998 09 2016-1 2011-12-7ab4bf4878d8f7661dfc20e9b8e18011 foosdfpfkusacimwkcsosbicdxkicaua 45097 02 2010-04-2016-0 0 01495c955be7ec5e7f3203406785aae0 foosdfpfkusacimwkcsosbicdxkicaua 29552 1260 21 10 rows × 41 columns #Feature Selection In [26]: features = Dataset[['channel_sales', 'cons_12m', 'cons_gas_12m', 'cons_last_month', 'forecast_cons_12m' 'forecast cons year', 'forecast_discount_energy', 'forecast_meter_rent_12m', 'forecast_price_energy_p1', 'foreca st_price_energy_p2', 'forecast price pow p1', 'has gas', 'imp cons', 'margin gross pow ele', 'margin net pow el 'net_margin', 'num_years_antig', 'origin_up', 'pow_max', 'price_p1_var', 'price_p2_var', 'price_p1_fix', 'price_p2_fix', 'price_p3_fix', 'num_years_renewal', 'avg_month_elec', 'av g_month_gas']] target = Dataset['churn'] In [27]: | #Creating dummies for categorical data Features = pd.get dummies(features) In [29]: #Splitting dataset to test and train from sklearn.model_selection import train_test_split x_train, x_test, y_train, y_test = train_test_split(Features, target, test_size=0.30, random_state=4, s tratify=target) **Data Modelling** Model 1 - Logistic Regression In [40]: | lr = LogisticRegression(max iter = 100000, dual = False) start = time.time() lr model = lr.fit(x train, y train) end = time.time() fit_time = (end - start) start = time.time() y_pred = lr_model.predict(x_test) end = time.time() pred time = (end - start) precision, recall, fscore, train_support = score(y_test, y_pred, pos_label= 1, average='binary') print('Fit time: {} / Predict time: {} ---- Precision: {} / Recall: {} / Accuracy: {}'.format(round(fit_time, 3), round(pred_time, 3), round(precision, 3), round(recall, 3), round((y_pred==y_te st).sum()/len(y_pred), 3))) Fit time: 1.164 / Predict time: 0.007 ---- Precision: 0.0 / Recall: 0.0 / Accuracy: 0.895 poor model. No precision and recall. Model 2 - Random Forest Classifier In [43]: | rf = RandomForestClassifier(n estimators=200, max depth=None, n jobs=-1) start = time.time() rf model = rf.fit(x_train, y_train) end = time.time() fit time = (end - start) start = time.time() y pred2 = rf model.predict(x test) end = time.time() pred_time = (end - start) precision, recall, fscore, train_support = score(y_test, y_pred2, pos_label= 1, average='binary') print('Fit time: {} / Predict time: {} ---- Precision: {} / Recall: {} / Accuracy: {}'.format(round(fit_time, 3), round(pred_time, 3), round(precision, 3), round(recall, 3), round((y_pred2==y_t est).sum()/len(y pred2), 3)))Fit time: 2.549 / Predict time: 0.115 ---- Precision: 0.698 / Recall: 0.063 / Accuracy: 0.904 Moderate precision and better accuracy, but low recall. Model 3 - Gradient Boost Classifier In [45]: gb = GradientBoostingClassifier(n estimators=150, max depth=11) start = time.time() gb_model = gb.fit(x_train, y_train) end = time.time() fit time = (end - start) start = time.time() y pred3 = gb model.predict(x test) end = time.time() pred time = (end - start) precision, recall, fscore, train_support = score(y_test, y_pred3, pos_label= 1.0, average='binary') print('Fit time: {} / Predict time: {} ---- Precision: {} / Recall: {} / Accuracy: {}'.format(round(fit_time, 3), round(pred_time, 3), round(precision, 3), round(recall, 3), round((y_pred3==y_t est).sum()/len(y_pred3), 3))) Fit time: 30.353 / Predict time: 0.101 ---- Precision: 0.588 / Recall: 0.104 / Accuracy: 0.904 Moderate precision and accuracy, but recall is still low. Model 4 - Stacking Classifier



