Project description

The following is a fictional case study designed to loosely resemble the work you might undertake on a future project. It will test your ability to handle big data and perform statistical/machine learning analyses as well as your ability to communicate your findings and derive commercial insight from your technical work.
You may perform the analyses using any computational language you wish (including at least one tool different from excel, since the majority of data sets we receive from clients are too large for us to be able to use it). Please submit your code along with your presentation and the requested results file by the date agreed with Gamma recruitment team.

Scenario:
Our client is a major utility company providing gas and electricity to corporate, SME and residential customers. In recent years, post-liberalization of the energy market in US, has had a growing problem with increasing customer defections above industry average. Thus, the client has asked us to work alongside them to identify the drivers of this problem and to devise and implement a strategy to counter it. The churn issue is most acute in the SME division and thus they want it to be the first priority.
The head of the SME division has asked whether it is possible to predict the customers which are most likely to churn so that they can trial a range of pre-emptive actions. He has a hypothesis that clients are switching to cheaper providers so the first action to be trialed will be to offer customers with high propensity of churning a 20% discount.

Your task:
We have scheduled a meeting in one week's time with the head of the SME division in which you will present our findings of the churn issue and your recommendations on how to address it.
You are in charge of building the model and of suggesting which commercial actions should be taken as a result of the model's outcome.
The first stage is to establish the viability of such a model. For training your model you are provided with a dataset which includes features of SME customers in January 2016 as well as the information about whether or not they have churned by March 2016. In addition to that you have received the prices from 2015 for these customers.
Of particular interest for the client is how you frame the problem for training. Given that this is the first time the client is resorting to predictive modelling, it is beneficial to leverage descriptive statistics and visualisation for extracting interesting insights from the provided data before diving into the model. Also while it is not mandatory, you are encouraged to test multiple algorithms. If you do so it will helpful to describe the tested algorithms in a simple manner.

Using the trained model you shall “score” customers in the verification data set (provided in the eponymous file) and put them in descending order of the propensity to churn. You should also classify these customers into two classes: those which you predict to churn are to be labelled "1" and the remaining customers should be labelled "0" in the result template.
You will submit this file with your presentation and your predictions will be scored with area under the ROC curve.
Finally, the client would like to have a view on whether the 20% discount offer to customers predicted to be churned is a good measure. Given that it is a steep discount bringing their price lower than all competitors we can assume for now that everyone who is offered will accept it. According to regulations they cannot raise the price of someone within a year if they accept the discount. Therefore offering it excessively is going to hit revenues hard.

Table 1 describes all the data fields which are found in the data. You will notice that the contents of some fields are meaningless text strings. This is due to "hashing" of text fields for data privacy. While their commercial interpretation is lost as a result of the hashing, they may still have predictive power.

Field name	Description
id	contact id
activity_new	category of the company's activity
campaign_disc_ele	code of the electricity campaign the customer last subscribed to
channel_sales	code of the sales channel
cons_12m	electricity consumption of the past 12 months
cons_gas_12m	gas consumption of the past 12 months
cons_last_month	electricity consumption of the last month
date_activ	date of activation of the contract
date_end	registered date of the end of the contract
date_first_activ	date of first contract of the client
date_modif_prod	date of last modification of the product
date_renewal	date of the next contract renewal
forecast_base_bill_ele	forecasted electricity bill baseline for next month
forecast_base_bill_year	forecasted electricity bill baseline for calendar year
forecast_bill_12m	forecasted electricity bill baseline for 12 months
forecast_cons	forecasted electricity consumption for next month
forecast_cons_12m	forecasted electricity consumption for next 12 months
forecast_cons_year	forecasted electricity consumption for next calendar year
forecast_discount_energy	forecasted value of current discount
forecast_meter_rent_12m	forecasted bill of meter rental for the next 12 months
forecast_price_energy_p1	forecasted energy price for 1st period
forecast_price_energy_p2	forecasted energy price for 2nd period
forecast_price_pow_p1	forecasted power price for 1st period
has_gas	indicated if client is also a gas client
imp_cons	current paid consumption
margin_gross_pow_ele	gross margin on power subscription
margin_net_pow_ele	net margin on power subscription
nb_prod_act	number of active products and services
net_margin	total net margin
num_years_antig	antiquity of the client (in number of years)
origin_up	code of the electricity campaign the customer first subscribed to
pow_max	subscribed power
price_date	reference date
price_p1_var	price of energy for the 1st period
price_p2_var	price of energy for the 2nd period
price_p3_var	price of energy for the 3rd period
price_p1_fix	price of power for the 1st period
price_p2_fix	price of power for the 2nd period
price_p3_fix	price of power for the 3rd period
churned	has the client churned over the next 3 months