**Project Problem statement Description**

Client is a major gas and electricity utility • Supplies to corporate, SME and residential customers Significant churn problem • Driven by power-liberalization of the energy market in Europe • Problem is largest in the SME segment

CLIENT HYPOTHESIS

It is possible to predict customers likely to churn using a predictive model Hypothesis that churn is driven by customer price sensitivity Client wants to try discounting strategy • SME division head suggests that offering customers at high propensity to churn a 20% discount might be effective

project team thinks that building a churn model to understand whether price sensitivity is the largest driver of churn has potential. The client has sent over some data and the client wants you to perform some exploratory data analysis.

The data that was sent over includes:

* Historical customer data: Customer data such as usage, sign up date, forecasted usage etc
* Historical pricing data: variable and fixed pricing data etc
* Churn indicator: whether each customer has churned or not
* **Sub-Task 1:**
* Perform some exploratory data analysis. Look into the data types, data statistics, specific parameters, and variable distributions. This first subtask is for you to gain a holistic understanding of the dataset.
* **Sub-Task 2:**
* Verify the hypothesis of price sensitivity being to some extent correlated with churn. It is up to you to define price sensitivity and calculate it.
* **Sub-Task 3:**
* Prepare a half-page summary or slide of key findings and add some suggestions for data augmentation – which other sources of data should the client provide you with and which open source datasets might be useful?
* For your final deliverable, please submit your analysis (in the form of a jupyter notebook, code script or PDF) as well as your half-page summary document.

**client\_data.csv**

* id = client company identifier
* activity\_new = category of the company’s activity
* 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\_modif\_prod = date of the last modification of the product
* date\_renewal = date of the next contract renewal
* forecast\_cons\_12m = forecasted electricity consumption for next 12 months
* forecast\_cons\_year = forecasted electricity consumption for the next calendar year
* forecast\_discount\_energy = forecasted value of current discount
* forecast\_meter\_rent\_12m = forecasted bill of meter rental for the next 2 months
* forecast\_price\_energy\_off\_peak = forecasted energy price for 1st period (off peak)
* forecast\_price\_energy\_peak = forecasted energy price for 2nd period (peak)
* forecast\_price\_pow\_off\_peak = forecasted power price for 1st period (off peak)
* 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
* churn = has the client churned over the next 3 months

**price\_data.csv**

* id = client company identifier
* price\_date = reference date
* price\_off\_peak\_var = price of energy for the 1st period (off peak)
* price\_peak\_var = price of energy for the 2nd period (peak)
* price\_mid\_peak\_var = price of energy for the 3rd period (mid peak)
* price\_off\_peak\_fix = price of power for the 1st period (off peak)
* price\_peak\_fix = price of power for the 2nd period (peak)
* price\_mid\_peak\_fix = price of power for the 3rd period (mid peak)

Note: some fields are hashed text strings. This preserves the privacy of the original data but the commercial meaning is retained and so they may have predictive power

Task2

The team now has a good understanding of the data and feels confident to use the data to further understand the business problem. The team now needs to brainstorm and build out features to uncover signals in the data that could inform the churn model.

Feature engineering is one of the keys to unlocking predictive insight through mathematical modeling. Based on the data that is available and was cleaned, identify what you think could be drivers of churn for our client and build those features to later use in your model.

First focus on building on top of the feature that your colleague has already investigated: **“the difference between off-peak prices in December and January the preceding year”**. After this, if you have time, feel free to get creative with making any other features that you feel are worthwhile.

Once you have a set of features, you must train a Random Forest classifier to predict customer churn and evaluate the performance of the model with suitable evaluation metrics. Be rigorous with your approach and give full justification for any decisions made by yourself as the intern data scientist.

Recall that the hypotheses under consideration is that churn is driven by the customers’ price sensitivities and that it would be possible to predict customers likely to churn using a predictive model.

If you’re eager to go the extra mile for the client, when you have a trained predictive model, remember to investigate the client’s proposed discounting strategy, with the head of the SME division suggesting that offering customers at high propensity to churn a 20% discount might be effective.

Build your models and test them while keeping in mind you would need data to prove/disprove the hypotheses, as well as to test the effect of a 20% discount on customers at high propensity to churn.

Task3

 some work on engineering the features within the cleaned dataset has been done and has calculated a feature which seems to have predictive power.

This feature is **“the difference between off-peak prices in December and January the preceding year”**.

Run the cells in the notebook provided (named feature\_engineering.ipynb) to re-create this feature. then try to think of ways to improve the feature’s predictive power and elaborate why you made those choices.

You should spend 1 - 1.5 hours on this. Be sure to make use of the “feature\_engineering.ipynb” notebook to get started with re-creating your colleagues' features.

**Sub-Task 2**

Now that you have a dataset of cleaned and engineered features, it is time to build a predictive model to see how well these features are able to predict a customer churning. It is your task to train a Random Forest classifier and to evaluate the results in an appropriate manner. We would also like you to document the advantages and disadvantages of using a Random Forest for this use case. It is up to you how to fulfill this task, but you may want to use the below points to guide your work:

* Ensure you’re able to explain the performance of your model, where did the model underperform?
* Why did you choose the evaluation metrics that you used? Please elaborate on your choices.
* Document the advantages and disadvantages of using the Random Forest for this use case.
* Do you think that the model performance is satisfactory? Give justification for your answer.
* (Bonus) - Relate the model performance to the client's financial performance with the introduction of the discount proposition. How much money could a client save with the use of the model? What assumptions did you make to come to this conclusion?

. When it comes to model evaluation and the explanation of your results, feel free to use the additional links below.

**If you are stuck:**

Sub-Task 1

* Think of ways to evaluate a feature against a label.
* Think of ways to add new features which would complement the already existing ones.
* Think of feature granularity.
* Remove unnecessary features.

Sub-Task 2

* Is this problem best represented as classification or regression?
* What kind of model performance do you think is appropriate?
* Most importantly how would you measure such a performance?
* How would you tie business metrics such as profits or savings to the model performance?