Task 3: Feature Engineering and Modelling

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.

Sub-Task 1

Your colleague has done some work on engineering the features within the cleaned dataset 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?

You should spend 1 - 1.5 hours on this. 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?

Estimated time for task completion: 2-3 hours depending on your learning style.