TASK BRIEF

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 fulfil 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.

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.

Making graphs in Python:
1. Matplotlib - Visualization and Python - https://matplotlib.org/
2. Seaborn - Statistical Data Visualization -
https://seaborn.pydata.org/index.html

Making graphs in R :
1. Create Elegant Data Visualizations Using the Grammar of Graphics -
https://ggplot2.tidyverse.org/
2. Plotly R Graphing Library - https://plotly.com/r/