From all results above, we can conclude customers were not sensitive to prices.

Sub-Task 3

Key findings

1. About 9.72% of customers changed providers.

2. Numeric variables on consumption are highly skewed.

3. Changes in prices does not affect customer churn.

Suggestions

1. Competitor price data – If other providers can give a much better offer than the current provider, customers were more likely to transfer to new provider even if their current prices dropped.

2. Need to clarify what values of zero in price data represent. If the prices of zero stand for free power or energy, what is the reason for that.

3. Other possible factors, such as customer satisfaction. For example, if providers could keep providing excellent customer services, it is very likely a rise in prices will not lead to customer churn.

Feature Engineering & Modelling

Background information

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.

Task

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 featured_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.

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.

Sub-Task

According to the task requirement, build a Random Forest model.

As there is no testing set, the "out-of-fold" cross validation strategy is used to evaluate the model's performance. Besides, the training set is a very imbalanced dataset, so I use precision and recall as the metrics, instead of accuracy (results of accuracy are still shown, but not for evaluating the model's performance). As the goal of this task is not finding the optimal parameters, I do not tune the model's parameters.

Based on the results above, we can see the performance is very bad. Although the accuracy is up to 90%, it is misleading and pointless, as we only focus on correctly predicting the positive class rather than the negative class. The model can only find out 6% of real positive samples. But luckily, in the predicted positive samples, up to 86% of them are real positive samples.

Bonus task: identify the impact of a 20% discount

This is a tricky problem.

- first, intuitively, we have concluded that changes in prices do not have a significant impact on customer churn. So, a discount should be unable to prevent customer churn.
- second, we cannot really identify whether a discount can prevent customer churn, as we do not exactly know what will happen in the future.

Therefore, I try to calculate the expected profit based on the probability of customer churn and evaluate the impact of the discount.

We need to predict the probability of churn when applying a 20% discount. Note: I still use the data on this year, not the next year.

Expected profit from only 349 customers (2%) rise after being offered discounted prices.
Among the 104 predicted churning customers, after being offered discounted prices,

- the expected profit from 61 customers increases and total expected profit increase by 2250001.
- the expected profit from 43 customers decreases and total expected profit decrease by 1138753.

So, it is expected that offering a discount to the predicted churning customers can bring extra profit of 1111248.

Based on the results above, it seems offering a discount to the predicted churning customers can increase the expected profit from these customers.