A. Business Understanding & Hypothesis Framing

Background Information Task

PowerCo is a major gas and electricity utility that supplies to corporate, SME (Small & Medium Enterprise), and residential customers. The power-liberalization of the energy market in Europe has led to significant customer churn, especially in the SME segment. They have partnered with BCG to help diagnose the source of churning SME customers.

A fair hypothesis is that price changes affect customer churn. Therefore, it is helpful to know which customers are more (or less) likely to churn at their current price, for which a good predictive model could be useful.

Moreover, for those customers that are at risk of churning, a discount might incentivize them to stay with our client. The head of the SME division is considering a 20% discount that is considered large enough to dissuade almost anyone from churning (especially those for whom price is the primary concern).

The Associate Director (AD) held an initial team meeting to discuss various hypotheses, including churn due to price sensitivity. After discussion with your team, you have been asked to go deeper on the hypothesis that the churn is driven by the customers' price sensitivities.

Your AD wants an email with your thoughts on how the team should go about testing this hypothesis.

The client plans to use the predictive model on the 1st working day of every month to indicate to which customers the 20% discount should be offered.

Task

You must formulate the hypothesis as a data science problem and lay out the major steps needed to test this hypothesis. Communicate your thoughts and findings in an email to your AD, focusing on the data that you would need from the client and the analytical models you would use to test such a hypothesis.

Solutions

Hi [AD],

In order to test the hypothesis of whether churn is driven by the customers' price sensitivity, we would need to model churn probabilities of customers, and derive the effect of prices on churn rates. We would need the following data to be able to build the models.

Data needed:

1. Customer data - which should include characteristics of each client, for example, industry, historical electricity consumption, date joined as customer etc.
2. Churn data - which should indicate if customer has churned
3. Historical price data – which should indicate the prices the client charges to each customer for both electricity and gas at granular time intervals

Once we have the data, the **work plan would be**:

1. We would need to define what price sensitivity is and calculate it
2. We would need to engineer features based on the data that we obtain, and build a binary classification model (e.g. Logistic Regression, Random Forest, Gradient Boosted Machines to name a few),
3. The best model would be picked based on the tradeoff between the complexity, the explainability, and the accuracy of the models.
4. We would subsequently dive deeper into why and how price changes impact churn.
5. Last but not least, the model would allow us to size the business impact of the client's proposed discounting strategy.

Regards, [Aditya Agral Serhansyah]

B. Exploratory Data Analysis (EDA)

Background Information Task

The BCG 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 AD 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

Task

1. **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.
2. **Sub-Task 2**: Verify the hypothesis of price sensitivity being to some extent correlated with churn.
3. **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?

EDA Summary

Findings

- Approximately 10% of customers have churned
- Consumption data is highly skewed and must be treated before modeling
- There are outliers present in the data and these must be treated before modeling
- Price sensitivity has a low correlation with churn
- Feature engineering will be vital, especially if we are to increase the predictive power of price sensitivity

Suggestions

- Competitor price data - perhaps a client is more likely to churn if a competitor has a good offer available?
- Average Utilities prices across the country - if PowerCo's prices are way above or below the country average, will a client be likely to churn?
- Client feedback - a track record of any complaints, calls or feedback provided by the client to PowerCo might reveal if a client is likely to churn

C. Feature Engineering & Modelling

Background Information Task

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

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?

Summary

The model has been built using the Random Forest Classifier showing the model evaluation results as follows:

Model Evaluation	Value
True positives	18
False positives	4
True negatives	3282
False negatives	348
Accuracy	0.9036
Precision	0.8181
Recall	0.0492

- Within the test set about 10% of the rows are churners (churn = 1).
- Looking at the true negatives, we have 3282 out of 3286. This means that out of all the negative cases (churn = 0), we predicted 3282 as negative (hence the name True negative). This is great!
- Looking at the false negatives, this is where we have predicted a client to not churn (churn = 0) when in fact they did churn (churn = 1). This number is quite high at 348, we want to get the false negatives to as close to 0 as we can, so this would need to be addressed when improving the model.
- Looking at false positives, this is where we have predicted a client to churn when they actually did not churn. For this value we can see there are 4 cases, which is great!
- With the true positives, we can see that in total we have 366 clients that churned in the test dataset. However, we are only able to correctly identify 18 of those 366, which is very poor.
- Looking at the accuracy score, this is very misleading! Hence the use of precision and recall is important. The accuracy score is high, but it does not tell us the whole story.
- Looking at the precision score, this shows us a score of 0.82 which is not bad, but could be improved.
- However, the recall shows us that the classifier has a very poor ability to identify positive samples. This would be the main concern for improving this model!

From this feature importance chart, we can observe the following points:

- Net margin and consumption over 12 months is a top driver for churn in this model
- Margin on power subscription also is an influential driver
- Time seems to be an influential factor, especially the number of months they have been active, their tenure and the number of months since they updated their contract

- The feature that our colleague recommended is in the top half in terms of how influential it is and some of the features built off the back of this actually outperform it
- Our price sensitivity features are scattered around but are not the main driver for a customer churning

In the strategy suggested by the SME division head we offer a 20% discount to all customers targeted. However, this might not be optimal either. We assumed before that customers offered a discount will not churn. However, that may not be true in reality. The discount may not be large enough to prevent churn.

In fact, we can predict the churn probability for each customer as a function of price, margin and other factors. Therefore, we can try to find a strategy for each customer that optimizes either their expected revenue or profit.

In order to go further, we'll need to try to:

- Change the level of discount offered overall
- Predict the response of customers to that discount (ie, the churn probability) based on how much that discount affects their prices, the revenue and margin.
  - Take care that we've applied the discount to all affected variables. To make this easier, we might want to retrain our model using a simpler set of variables where we know that we can factor the discount correctly into the predictors.
- Find the discount level that balances customer retention vs the cost of false positives.

In fact, this could be turned into a 2D optimisation problem:

- **Objective**: maximize net revenue (ie including the benefits of true positives and the cost of false positives)
- **Decision variables**:
  - Level of discount offered, and
  - Fraction of people who are offered a discount

An even more sophisticated strategy is to find the right level of discount for each customer that maximizes their predicted revenue or margin.

D. Findings & Recommendations

Background Information Task

The AD wants you to draft an abstract (executive summary) of your findings so far.

Task

Develop an abstract slide synthesizing all the findings from the project so far, keeping in mind that this will be for the key stakeholders meeting which the Head of the SME division, as well as other various stakeholders, will be attending.

Summary

1. **Churn is indeed high** in the SME division
   - 9,7% across 14.606 customers
2. **Predictive model is able to predict churn** but the main driver is not customer price sensitivity
   - Yearly consumption, forecasted consumption and net margin are the 3 largest drivers
3. **Discount strategy of 20% is effective** but only if targeted appropriately
   - Offer discount to only to high -value customers with high churn probability