BCG

One of the most important tasks of a data scientist is to understand the business problem at hand and lay out an approach to address it.

Understand problem: translate into data science.

Do

* Determine the client data needed for analysis
* Outline the techniques you'll use to investigate your client's problem
* Write an email to your Associate Director summarizing your approach

Your team has been assigned a new project for a client called **PowerCo**

“DevOps is a software development practice that promotes collaboration between development and operations, resulting in faster and more reliable software delivery” (Microsoft.com).

What exactly is a Data Scientist?

* A Data Scientist works with data to try and deliver value to a business.
* But how is that achieved?
  + Data Science can sometimes overlap with domains such as Data Analysts, Data Engineers and DevOps.
  + How much the Data Scientists role overlaps with these other roles really depends on the company, but in essence a Data Scientist is generally focused on modeling data to be able to accurately predict an outcome, for example, predicting how likely customers are to leave.
  + **4 core skills are used as a Data Scientist: statistics, mathematics, programming and communication.**

**BCG GAMMA Methodology: (**business understanding & problem framing)

Key roles and responsibilities of a Data Scientist at BCG GAMMA

BCG GAMMA is transforming businesses using data science to help companies generate competitive advantage. To do this, we typically follow a 5-step methodology:

1. **Business understanding & problem framing:** what is the context of this problem and why are they trying to solve it?
2. **Exploratory data analysis & data cleaning:**what data are we working with, what does it look like and how can we make it better?
3. **Feature engineering:**can we enrich this dataset using our own expertise or third party information?
4. **Modeling and evaluation:**can we use this dataset to accurately make predictions? If so, are they reliable?
5. **Insights & Recommendations:**how we can communicate the value of these predictions by explaining them in a way that matters to the business?

The brief from PowerCo

The Associate Director (AD) of the Data Science team held a team meeting to discuss the client brief. You’ll be working closely with Estelle Altazin, a senior data scientist on your team.

Here are the key takeaways from the meeting:

* Your client is **PowerCo**- a major gas and electricity utility that supplies to small and medium sized enterprises.
* The energy market has had a lot of change in recent years and there are more options than ever for customers to choose from.
* PowerCo are concerned about their customers leaving for better offers from other energy providers. When a customer leaves to use another service provider, this is called **churn**.
* This is becoming a big issue for PowerCo and they have engaged BCG to help diagnose the reason why their customers are churning.

During the meeting your AD discussed some potential reasons for this churn, one being **how “sensitive” the price is. In other words, how much is price a factor in a customer’s choice to stay with or leave PowerCo?**

So, now it’s time for you to investigate this hypothesis.

AD: Associate Director.

COB: Close of Business, it means the end of the business day.

Your task - we need to understand PowerCo’s problem in detail

First things first, you and Estelle need to understand the problem that PowerCo is facing at a deeper level and plan how you’ll tackle it. If you recall the 5 steps in the Data Science methodology, this is called “business understanding & problem framing”.

Your AD wants you and Estelle to email him by COB today outlining:

1. the data that we’ll need from the client, and
2. the techniques we’ll use to investigate the issue.

**Use the text field below to write your email, here’s what you’ll need to include:**

You must formulate PowerCo’s issue as a problem using the 5 step data science process and lay out the major steps needed to test it.

1. What do you think are the key reasons for a customer deciding to stay with or switch energy providers? For example: price, is it clean energy, customer service, location etc.
2. What data do you think would be useful in order to investigate these key reasons? E.g. customer purchasing trends over past 5 years, location of business etc.
3. If you were to get this data, how could you analyse or visualize it to test whether these reasons may have an impact on churn?

Use the steps to lay out steps to test power co’s issue

* Key reasons?
* Useful data?
* Analyze & Visualize Impact

**My Email Answer**

Dear Associate Director,

In this email, Etelle and I mention the reasons we think a customer would leave, which helps us to point out the data we’ll need from PowerCo. We also describe the techniques that we will use for analyzing and visualizing the data.

Reasons

Key reasons we think customers would likely switch include information related to convenience, customer satisfaction, price, length of power outages, location and customer service.

Useful Data

In order to identify which attributes have the greatest impact on customer churn, we’ll need the data related to the following topics, subtopics and questions.

* Customer Satisfaction
  + Customer Ratings of PowerCo
  + Frequency and duration of power outages
* Customer Service
  + Benefits other companies provide that PowerCo does not
* General customer information
  + Date person/business became a customer
  + Spending or Utility bill history
  + Customer income data
  + Customer education levels
  + Whether the customer is residential, commercial or manufacturer
  + Location
* Additional information for customers who switched
  + Reason they left, if available
  + Date the customer left
  + Number of customers leaving the company each month
* Price history
  + History of price data, including factors determining what a customer is charged
  + Date and amounts of price hikes
  + What is the price comparison between PowerCo and other utility companies
* Costs for PowerCo
  + Fuel costs
  + Electricity costs
  + Does PowerCo generate their own gas and electricity using their own power plants and oil rigs, or does PowerCo purchase gas and electricity from other energy producers?

Analytical Techniques

First, we need to see the shape or distribution of each attribute to better understand what is going on. Also, for numerical attributes, descriptive statistics may be useful, such as mean, median, variance and standard deviation. Then, running simple correlation and regression statistics will be crucial for numerical attributes, such as price, to show the relationship those attributes have with customer churn. Those correlations can be effectively visualized with scatter plots or line charts. For categorical attributes, such as location or customer type, the relationship can be shown using percentages and proportions. For example, the percentage of the different types of customers who leave and when could prove insightful. The proportion of percentages can be effectively visualized using proportional bar charts. In addition, a map that uses color to show which areas have the highest customer churn would be useful.

We looking forward to presenting our insights to PowerCo to help them resolve this issue.

Thank you,

Mark Bahr

Junior Data Scientist

**Example Answer**

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.

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 need to define what price sensitivity is and calculate it
2. We need to prepare the data and engineer features
3. Then, we can test our hypothesis using a binary classification model (e.g. Logistic Regression, Random Forest, Gradient Boosted Machines to name a few)
4. We would choose a model from one of the tested algorithms based on the model complexity, the explainability, and the accuracy of the models.
5. With the trained model, we would be able to extrapolate the extent to which price sensitivity influences churn

Regards, [Your name]

**Task 2: Exploratory Data Analysis**

What You’ll Learn

* How to investigate whether price sensitivity is the most influential factor for a customer churning
* How to use frameworks to conduct exploratory data analysis

What You’ll Do

* Use python to analyze client data
* Create data visualizations to help you interpret key trends

Task: Get some initial insights, values and visualization. Descriptive statistics & vizualizing. Explore the data and send your code with comments on your key findings when you’re done.

Question: Is price sensitivity the most influential factor for customer churing, and, if not, to what extent does price sensitivity influence churn?

**What is price sensitivity?**

**The degree to which demand changes when the cost of a product or service changes. Some consumers won’t pay more if a lower-priced option is available.**

Price sensitivity is **the degree to which demand changes when the cost of a product or service changes.**

In the context of PowerCo, the “demand” refers to the demand for energy consumption.

Price sensitivity is commonly measured using the price elasticity of demand, which states that some consumers won't pay more if a lower-priced option is available.

**What is price elasticity of demand?**

Price elasticity of demand is a measurement of the change in consumption of a product in relation to a change in its price

Complete the quick knowledge check and move onto your exploratory data analysis.

2 data sets for exploratory analysis.

**What is exploratory data analysis?**

Exploratory data analysis (EDA) is a technique used by a Data Scientist to gain a holistic understanding of the data that they are working with.

It is mainly based around using statistical techniques (such as descriptive statistics) and visualizations to gain a deeper understanding of the statistical properties that the data holds.

The client has sent over 3 data sets (shown below):

1. Historical customer data: Customer data such as usage, sign up date, forecasted usage etc
2. Historical pricing data: variable and fixed pricing data etc
3. Churn indicator: whether each customer has churned or not

You need to analyze the following using Python:

* The data types of each column
* Descriptive statistics of the dataset
* Distributions of columns

Estelle has provided a starter Jupyter notebook has been provided for you to use as a template to complete your work.

**Here are some tips to help you:**

Let’s take a look at the 2 data sets PowerCo. has sent over:

* A first good step is to review the data to make sense of the columns…
  + **Hint:** Look *at data types of column to gain a better understanding of what the columns mean. This is why data description documents are important - they describe exactly what the columns represent.*
* Once you understand the columns in the dataset. Now you want to look at how the values in the data vary…
  + ***Hint:****this is why reporting descriptive statistics is useful because it’ll tell you some basic statistical properties of the columns in the data. It will also tell you how many values feature within a column, e.g. does a column only have 1 unique value or 100? This is useful to know because you can then start to build a picture of what this data represents.*
* You now understand how the values vary and what the data represents - next up, it can be useful to visualize some of this…
  + ***Hint:****Not all visualizations are useful. Keep the visualizations simple and always keep in mind what you’re trying to show. E.g. if you want to see how the distribution of a column looks and that column has 1000 unique values, using a pie chart would not be good because it would become too crowded! If the values are numeric, a distribution plot would be more appropriate.*
  + ***Hint:****make sure to use the starter Jupyter notebook provided, as this will show you some example visualizations and sample code to use!*

At this stage, you should now have a clearer understanding of what the data is and how it looks. This framework is not exhaustive, but it shows how you could start to build your own framework for analysing data.

When you’re ready, upload your Python code notebook to complete this task.

‘cons\_12m’, ‘cons\_gas\_12m’, ‘cons\_last\_month’ forecast\_cons\_12m’ forecast\_cons\_year forecast\_discount\_energy forecast\_meter\_rent\_12m forecast\_price\_energy\_off\_peak forecast\_price\_energy\_peak forecast\_price\_pow\_off\_peak imp\_cons margin\_gross\_pow\_ele margin\_net\_pow\_ele nb\_prod\_act net\_margin num\_years\_antig pow\_max churn

1st per

* forecast\_price\_energy\_off\_peak = forecasted energy price for 1st period (off peak)
* forecast\_price\_pow\_off\_peak = forecasted power price for 1st period (off peak)
* price\_off\_peak\_var = price of energy for the 1st period (off peak)
* price\_off\_peak\_fix = price of power for the 1st period (off peak)

2nd per

* price\_peak\_var = price of energy for the 2nd period (peak)
* price\_peak\_fix = price of power for the 2nd period (peak)
* forecast\_price\_energy\_peak = forecasted energy price for 2nd period (peak)

3rd per

* price\_mid\_peak\_var = price of energy for the 3rd period (mid peak)
* price\_mid\_peak\_fix = price of power for the 3rd period (mid peak)

My notebook was submitted

**What did they say about their example notebook:**

**Explanation**

Getting set up - This task is focused on exploratory data analysis of the client and price data provided:

* The first thing you should do is download the provided Jupyter notebook and the CSV datasets.
* To run the notebook, you need to make sure that you provide the path for the CSV files so that you can load the data.
* By running the cells that exist within the notebook from Estelle, this will show you what the two datasets look like, it will provide you with code to produce descriptive statistics and it will also give some examples and sample code on how to visualize the data.

Analysis - Once you’ve run the cells provided, it was your job to build on this exploratory analysis:

* The visualization provided by Estelle shows how many companies churned vs. how many companies did not churn. We can see from this that the churn rate is approximately 10%. This is actually a very good churn rate, the closer the rate is to 0%, the better.
* The next series of visualizations were created in an attempt to try and dive deeper into how churn changes based on other factors (using other columns). This is useful for us to investigate because it may help us to understand factors that drive churn.
* In the notebook we visualize churn vs. sales channel, contract type, number of products, number of years and origin/contract offer.
* For example:
  + We see that for sales channel, there are some sales channels that yield customers churning but there are also other sales channels that have no customers churning.
  + For contract type, we see quite an even split for customers churning. This is interesting because this may suggest that contract type is not a driving factor towards churn rate.
* Additionally, for some columns their distributions with churn rate included. This is useful for us to understand because based on the distribution of a column, this could affect our feature engineering later.
* We look at the distribution of consumption, subscribed power and forecast in the notebook.
* For example:
  + We notice that the distribution of consumption is very skewed, this is called a positive skew since it is biased towards lower values on the x axis.
  + This is interesting because you may decide to treat this column to reduce the skewness later on during feature engineering. But also because we may want to visualize if there are any outliers within this column.
  + To investigate outliers, we use a boxplot. From the boxplot we can see that with the column as it is there are definitely some outliers. Once again this is interesting because we may choose to remove some of these outliers later.

**Task 3: Feature Engineering and Modeling**

## What you'll learn

* How feature engineering can be used to test hypotheses
* How to build features to analyse the data for PowerCo

## What you'll do

* Use Python to build a new feature for your analysis

Task: Brainstorm and create features that could be fed into the churn model.

* Feature based on the difference between off-peak prices in Decembar and January of the preceding year. Build on this feature by creating new features from it.
* What are the main drivers of churn for PowerCo based on the data that we have?
* Use your enginereed features to uncover insight from data by training a model to predict churn probabilities. (Engineeering and modeling)
  + Balance complexity with overall performance and reliability
  + Focus on refining a single model
  + Use a random forest
  + Evaluate the model
  + Measure against assessment criteria that makes sense

Make sure to give full explanations for which work and give justification for any choices you make. If you have time, investigate the feasibility of the discount strategy that was suggested for customers likely to churn.

What is feature engineering?

Feature engineering refers to:

* **Addition**
* **Deletion**
* **Combination**
* **Mutation**

of your data set to improve machine learning model training, leading to better performance and greater accuracy.

In context of this task, feature engineering refers to the engineering of the price and client data to create new columns that will help us to predict churn more accurately.

Effective feature engineering is based on sound knowledge of the business problem and the available data sources.

Here’s what you need to think about before you submit your work

**Your task is to create new features for your analysis and upload your completed python file.**We'll show you an example answer on the next step, but we encourage you to give it a go first!Below are some tips on how to get started.

As before, a good way to quickly learn how to effectively feature engineer is to build a framework to follow. Below is an example of how you could attempt this task:

**First - can we remove any of the columns in the datasets?**

* There will almost always be columns in a dataset that can be removed, perhaps because they are not relevant to the analysis, or they only have 1 unique value.

**Second - can we expand the datasets and use existing columns to create new features?**

* For example, if you have “date” columns, in their raw form they are not so useful. But if you were to extract month, day of month, day of year and year into individual columns, these could be more useful.

**Third - can we combine some columns together to create “better” columns?**

* How do we *define* a “better” column and how do we *know which* columns to combine?
  + We’re trying to accurately predict churn - so a “better” column could be a column that improves the accuracy of the model.
  + And which columns to combine? This can sometimes be a matter of experimenting until you find something useful, or you may notice that 2 columns share very similar information so you want to combine them.

**Finally - can we combine these datasets and if so, how?**

* To combine datasets, you need a column that features in both datasets that share the same values to join them on.

At this stage, your data could look vastly different, or may have just some subtle differences to how it was before.

You will be done with this task when you’re happy with the new set of features that you’ve created and you think you’re ready to build a predictive model to see which of these features are useful for predicting churn. Upload your python file and move onto the example answer.

Look for difference in prices between months. Find out where PowerCo has most its customers, find the temperature extremes, and take the diff between least and greatest. For example, diff between prices in April/May and January, if it’s a cold place, or price diff between March and August if a hot place. Perhaps those would make more of a difference than the dec & jan prices.

For each cust id:

Find the max, min and diff btwen max and min

Find the max val & add as a value tcolumn to the data frame

Find the min value

Id max min diff (merge with regular data)

Biggest Diff?

|  |  |  |
| --- | --- | --- |
| Variables | Higher | lower |
| 1 | 6.24 | 6.735 |
| 2 | 7.285 | 7.297 |
| 3 | 7.003 | 6.55 |
|  |  |  |
|  |  |  |
|  |  |  |