Task 4: Findings & Recommendations

What you'll learn

- How predictive modelling can be used to indicate churn risk
- How to communicate your insights with clients

What you'll do

- Build a predictive model for churn using a random forest technique
- Write an executive summary with your findings

We're now ready to begin predicting churn!

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.

Estelle has informed you that a classification model would be best for this task, and has suggested that you try the Random Forest classifier.

What is classification?

When you are trying to predict an outcome, the result that you're trying to predict can either be:

- A continuous number, e.g. an employees salary
- Or a discrete value, e.g. a job title

In our example, we are trying to predict whether or not a client will churn, so it will only ever been 1 of 2 values (True/False, 1/0, etc…).

If the outcome that you're trying to predict has a fixed number of discrete values, this is a classification problem, as you are trying to "classify" the observations in the data. If the outcome is a continuous number, this is a regression problem. We will not cover regression problems in this task.

And how does a Random Forest work?

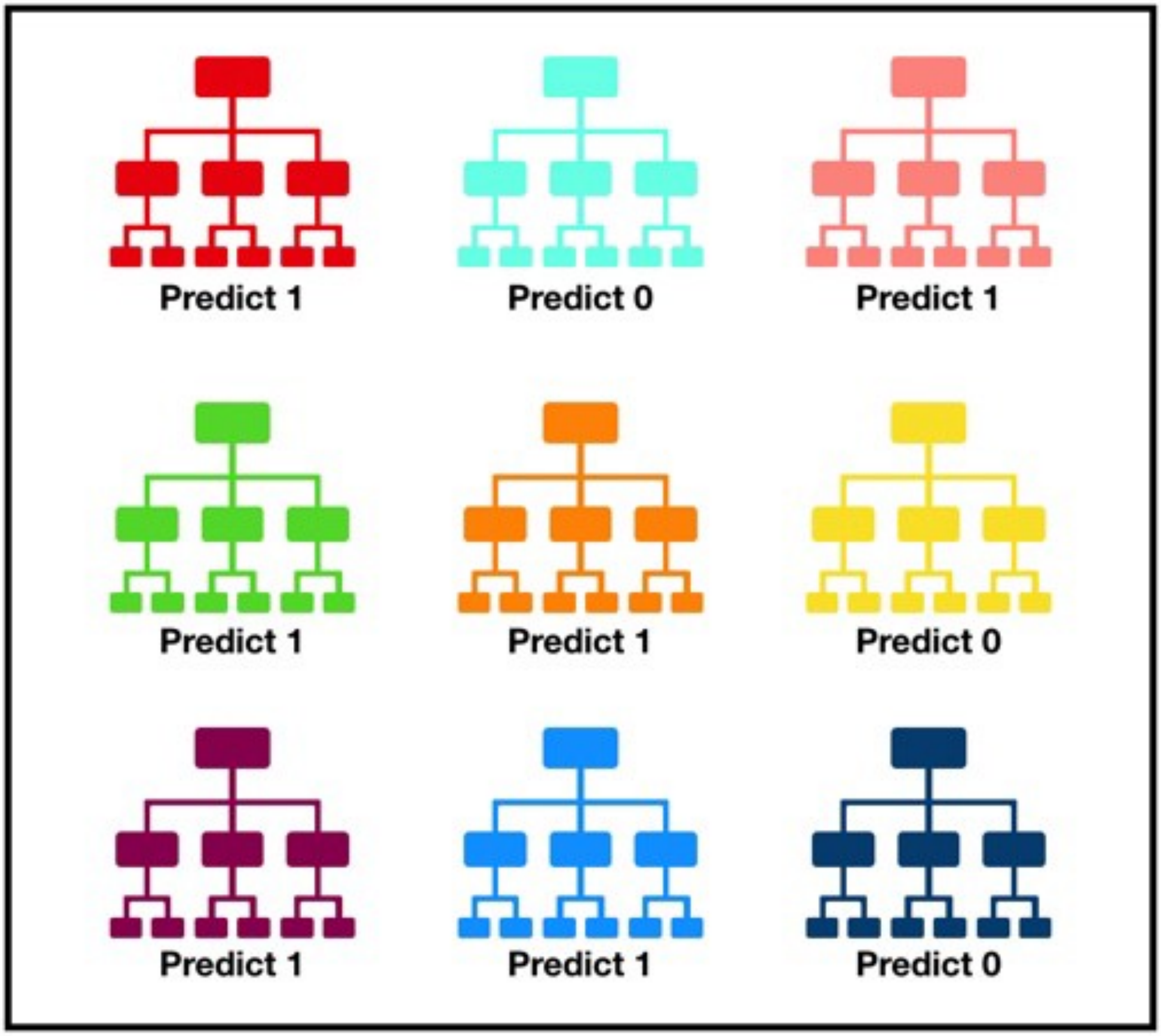
A random forest is a supervised learning algorithm which means that you must provide the algorithm with a set of features, as well as the outcome that you're trying to predict, in our case churn.

The way it makes predictions is by building a set of decision trees on different samples of the data and by taking a majority vote to decide what prediction to make.

To visualize this, the image below shows 9 decision trees and they are all trying to predict an outcome which is either a 1 or a 0 (similar to our case, where if someone has churned you see a 1, and if they haven't you see a 0).

The random forest would look at all the predictions generated from the 9 trees. You can see that 6 trees have predicted 1 and 3 have predicted 0. Therefore, the random forest would take the majority vote and present it's prediction as equal to 1.

If you wish to learn more on how this algorithm works, you can read more here.



Tally: Six 1s and Three 0s

Prediction: 1

Outline for making your predictions

It is your task to:

- Train a random forest classifier to predict churn
- Evaluate the predictions using evaluation metrics to demonstrate how accurately the model has performed

Estelle has provided a Jupyter notebook to get started. You should use this as a template to complete the work for this task.

Furthermore after the previous task of feature engineering, Estelle conducted a further review and has provided you with a final dataset to use this for this task, named “data_for_predictions.csv”. Be sure to use this dataset for this task.

You will notice that within the notebook, Estelle has imported various packages that would be used. One of them is named “scikit-learn”. This is an open source machine learning package and will be the source of the random forest model, as well as other things, that we use.

For more information on how to use the Random Forest classifier in scikit-learn, see the documentation site here

The outputs of your work will be shared with the AD and Estelle has given you a few points to include within the notebook:

- Why did you choose the evaluation metrics that you used? Please elaborate on your choices.
- Do you think that the model performance is satisfactory? Give justification for your answer.
- Make sure that your work is presented clearly with comments and explanations

Example Answer

Great work! Take a look at the example answer below to see how a professional would have attempted this task. Think about what you did well and how you can improve.

Explanation:

This final task is focused on building the predictive model using the CSV file that Estelle has shared.

- This CSV file contains a set of cleaned and engineered features so that you can focus purely on training your predictive model.
- You should download the Jupyter notebook and CSV file and run the cells provided in the notebook.
- These cells will load the data and create train and test samples of the data.

- It is important to split your data into train and test samples so then you can measure how well the trained model performs on an unseen set of data.
- This is a massively important thing to do when building a predictive model, otherwise you will have no way of measuring how well your model is able to predict churn for new customers!
- The code in the notebook provides you with skeleton code to create the random forest classifier, but it is your job to fill in the details of the code by using the documentation site provided.
- By adding in values for parameters within the random forest and by fitting the model on the training data, you will have a trained model to predict churn!

Now the most important part, evaluation of the model:

- It is left for you to decide how to evaluate the performance of the model. In general, you want to use metrics that reflect honestly how well the model has performed.
- In the notebook we use 3 metrics, accuracy, precision and recall.
- The reason why we are using these three metrics is because a simple accuracy measure (what percentage did I predict correctly) is not always a good measure to use.
- To give an example, let's say you're predicting heart failures with patients in a hospital and there were 100 patients out of 1000 that did have a heart failure.
- If you predicted 80 out of 100 (80%) of the patients that did have a heart failure correctly, you might think that you've done well! However, this also means that you predicted 20 wrong and what may the implications of predicting these remaining 20 patients wrong? Maybe they miss out on getting vital treatment to save their lives.
- As well as this, what about the impact of predicting negative cases as positive (people not having heart failure being predicted that they did), maybe a high number of false positives means that resources get used up on the wrong people and a lot of time is wasted when they could have been helping the real heart failure sufferers.
- This is just an example, but it illustrates why other performance metrics are necessary such as precision and recall, which are good measures to use in a classification scenario like this.

- After calculating the 3 metrics, we can see that we're able to accurately identify clients that do not churn, but not so accurately identify clients that will churn. Our model is predicting a high percentage of clients to not churn, when in fact they did!
- This tells me that the current set of columns are not a good set of features to predict churn. As the data scientist, it would normally be my job to go back and try to engineer a set of features that is able to predict churn more accurately.

Finally, we produce a feature importance chart to visualise which features were indeed useful within the model and which ones weren't.

- We can see that net margin and consumption over 12 months were important, to name a few.
- However the price sensitivity features are scattered around and do not shine through as a main driver for churn in their current form.

Finally, let's create a quick summary for the client

Before we finish up, the client wants a quick update on the project progress. Your AD wants you to draft an abstract (executive summary) of your findings so far.

Here is your 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.

Note: a steering committee meeting is a meeting where the BCG team presents key findings and recommendations (and/or project progress) to key client stakeholders.

Please use the template below and submit your summary slide in PDF format. We'll show you an example answer on the next step

A few things to think about for this abstract include:

- What is the most important number or metric to share with the client?
- What impact would the model have on the client's bottom line?

Please note, there are multiple ways to approach the task and that the sample answer is just one way to do it.

If you are stuck:

- What do you think the client wants to hear? How much detail should you go into, especially with the technical details of your work?
- Always test what you write with the “so what?” test, i.e. sharing a fact, even an interesting one, only matters if the client can actually do something useful with it. E.g. 60% of your customers are from City A is pointless, but customers in City A should be prioritized for giving discount as they are among your most valuable ones, if true, is an actionable finding.