FINDINGS AND RECOMMENDATIONS

4.1 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.

4.2 WHAT IS CLASSIFICATION?

When you are trying to predict an outcome, the result that you are 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 or 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. This task will not cover regression problems.

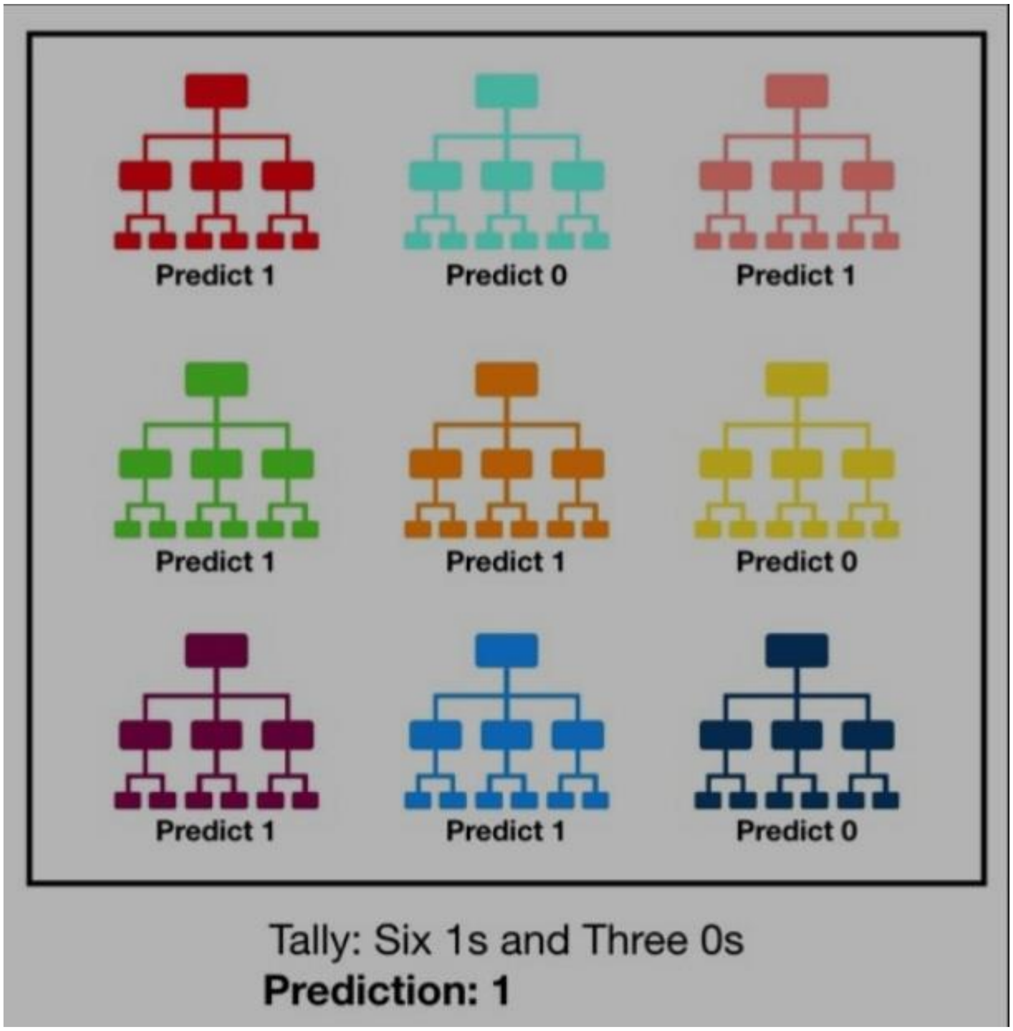
4.3 AND HOW DOES THE RANDOM FOREST WORK?

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 chase churn.

The way it makes prediction 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.



4.4 QUESTION

Is our task of predicting churn classification or a regression problem?

A. [✓ Classification]

B. Regression

How does a random forest classifier make its predictions?

A. Chooses a random outcome

B. [✓ Takes the majority vote of its decision trees]

C. Takes the outcome of 1 decision tree

4.5 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

4.6 EXPLANATION

This final task is focused on building the predictive model using the CSV file

- 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!

- 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 predicted 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 patient wrong? Maybe they miss out on getting vital treatment to save their life.
  - 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!

Finally, we produce a feature importance chart to visualize 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.