Summary Report: Decision Tree Classification Model

**Dashboard**

Model Evaluation: within the model evaluation tab in Power BI, you will be able to view the performance metrics of the decision tree model, it contains four visualisations

Classification Report – this visualisation displays the details around the model’s accuracy and recall.

**Confusion Matrix -** This visualises the model’s predictions, predicted Positives, Predicted negatives against the actual results of the recent campaign.

**Feature Importance –** This displays the parameter used by the decision tree to make it predictions and orders them in terms of importance.

**Receiver Operating Characteristic (ROC) Curve -** that shows how well a test can separate two things by plotting the rate of correct identifications against the rate of incorrect identifications.

**Decision Tree Tab –** this shows the routes and questions the decision tree used to get to its predictions, if you zoom in you can see each question and the split of customers at each point.

**Feature Selection:**

To train the model I kept all data fields except duration and campaign because this was a field that would only be known after the campaign, the benefit of using decision trees is that you can retain all the features and the tree will decide which ones are best to make the predictions the feature importance visual displays the best features to use.

**Advantages & disadvantages of using Python & Power BI**

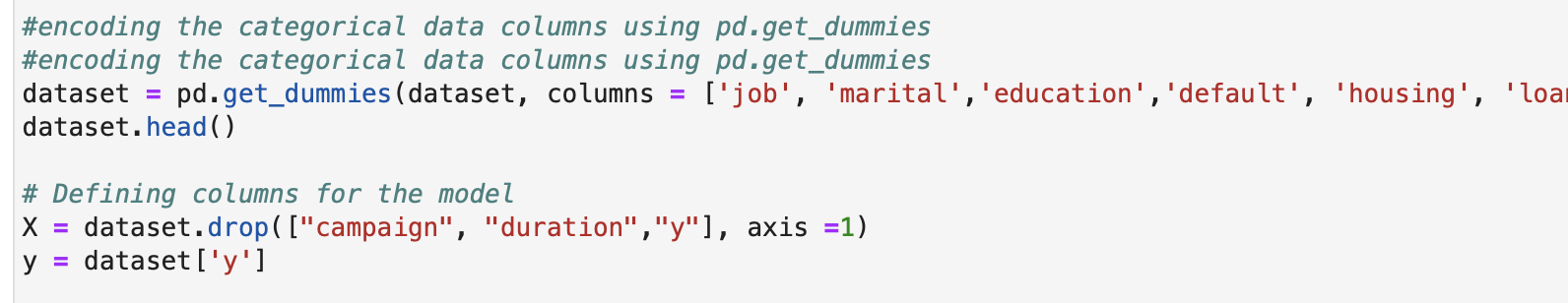
Utilising Python to build your model is very useful you can easily manipulate the data into the format needed to build the predictive model then in power BI you can visualise this for the end user so they just see the outputs of the model, which is better for non-technical colleagues who might not need to understand the coding side of your project, the drawback is that not everything is fully compatible with power BI, and it can be quite challenging adapting your code from python to work correctly in power BI but is achievable.

**Challenges**

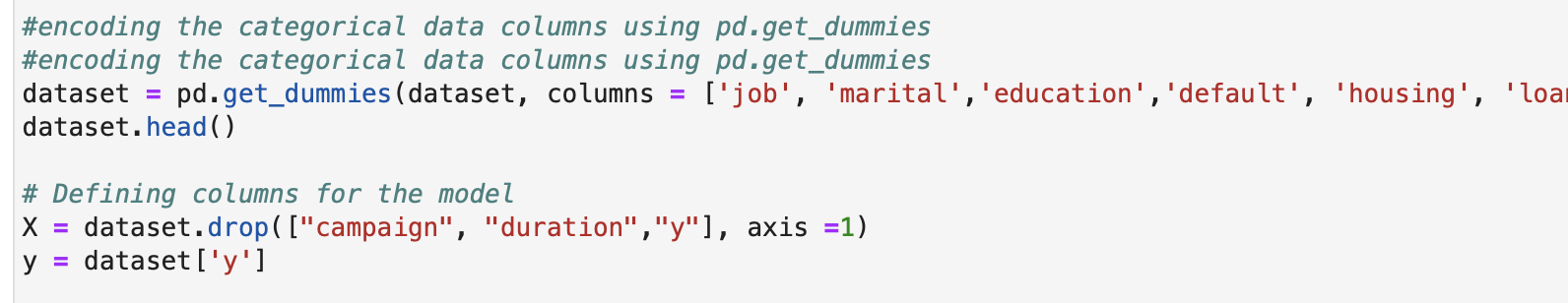
There were a few challenges with integrating the visualisations in power BI mainly regarding the Confusion Matrix and Power BI switching the axis, also an issue which required me to upgrade the version of seaborn that was needed to visualise that correctly which took some time, also the classification report visualisation was difficult to get the layout to match correctly with my python visual

**Development Process**

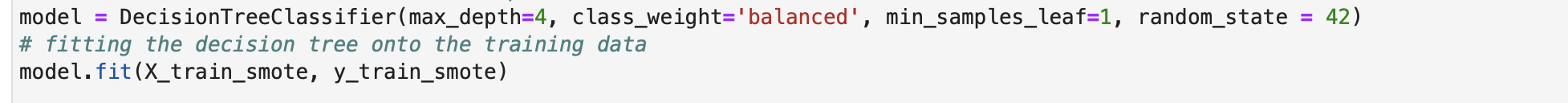
To build the model, I have first removed the duration & campaign columns from the dataset, I’ve then used one hot encoding to transform the categorical data into Boolean fields so that the decision tree can use this information,



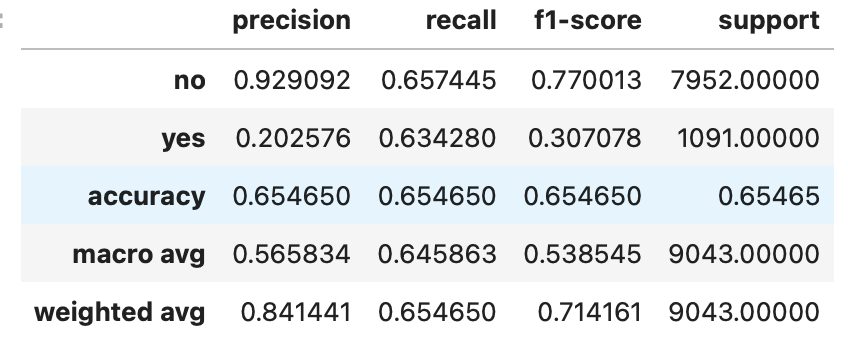
I’ve then used SMOTE (Synthetic Minority Over-sampling Technique) with is an oversampling technique used to it creates new, synthetic examples that are similar to the existing ones in the smaller group, helping to balance out the numbers between different groups in the data.



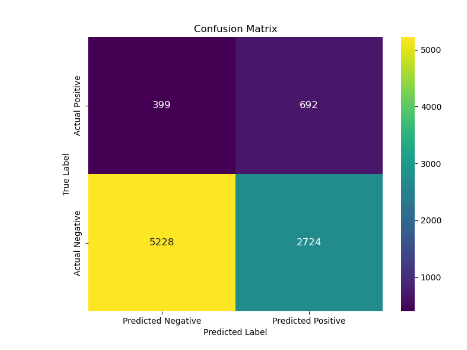
I’ve then used a decision tree classification model with a max depth of 4 questions, and a class weight of balanced meaning it treats the outcomes as equal.



This resulted in a model with the following metrics



The metrics highlights that the model is 20% accurate at predicting whether the customers will subscribe to a fixed term deposit, this is low overall, but the recall is 0.63 which shows that the model was able to predicts 63% of the customer that subscribed within the Yes predictions. The reason for the overall low accuracy of the model for predicting yes was that the model predicts 2724 that where positive when they didn’t subscribe. The reason I’ve decided to go with a model that has low accuracy but higher recall is that you have more potential to gain subscribers, with the higher accuracy model for predicting yes it only predicted around 200 yes subscriptions correctly whereas the low accuracy high recall predicted 692 yes correctly overall although you would be calling more customers overall you would end up with a lot more customers subscribing then with the higher accuracy model.



Future recommendations would be to gain more information about our customer base so we can train the model further or look into other types of classification model to see whether we can build a better predictive model one to look into further would be to create a neural network classification model.