Lloyds Banking Group

Data Analyst Incubation Program

Sprint 1 : Direct Marketing Campaign – Decision Tree Report

Contents

Introduction

As you already know, the marketing team are concerned with determining who amongst their customers are receptive to the Term deposit financial product campaigns. Currently the only way they can determine which clients to approach is to simply identify customers who don’t already have a Term deposit.

Term Deposit: In retail banking, it’s considered a financial product, a type of savings account where money is deposited for a fixed period, known as the term. It usually offers a higher interest rate compared to regular savings accounts.

Business need

Customers have expressed frustration around the high frequency of marketing campaigns and the lack of relevance to their needs. Our marketing division have requested that we develop a predictive model that will inform targeted marketing campaigns. The marketing department have also requested we rank informative data points (features) about customers to see which of these contribute most to the predicted outcome.

Objective

The Project Manager has scoped the project request and determined that a decision tree predictive model built using Python can be scripted into a Live Report using Power BI Desktop.

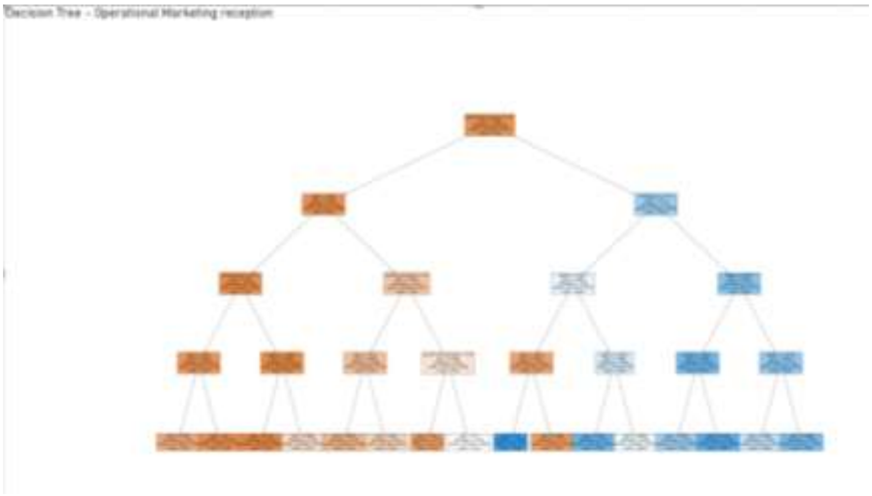
The dataset is taken from a direct marketing campaign for retail banking dataset `direct_marketing_dataset.csv` .
The resulting predictive model must be operational to the context; meaning it must only include data that is known before the customer is contacted about a term deposit. For example, the feature *Duration* measures the length of the call where the Term deposit is discussed but this data is known after the event. This is because the goal is to utilise features known about customers before the marketing call takes place

Communicating Results to stakeholders

Decision Tree Visual Report:

- Live Data Connection to dataset
- Decision Tree of max depth 4
- Each node displays the Sample Size and feature split

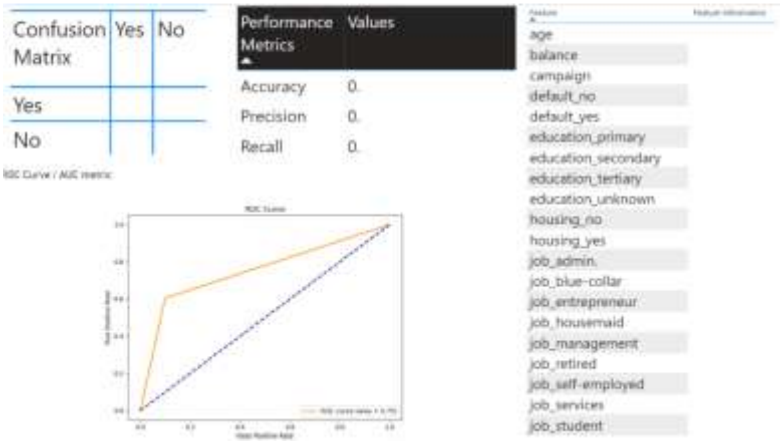
Must show entire tree and be legible – even if this requires magnification. Tip: In Power BI Dashboard view. Select Fill page or fit width. In Python, use matplotlib to alter **`fig(width, length)`** and fontsize, to achieve the desired legibility.



Evaluation and Features Visual Report:

Utilising the same dataset and modelling parameters, construct a PowerBI report with multiple components that displays the following:

- Most Informative features
- Performance Metrics
- (Advanced) Confusion Matrix
- (Advanced) ROC / AUC



Confusion Matrix	Yes	No
Yes		
No		

Performance Metrics	Values
Accuracy	0.
Precision	0.
Recall	0.

ROC Curve / AUC metric

Feature	Feature Information
age	
balance	
campaign	
default_no	
default_yes	
education_primary	
education_secondary	
education_tertiary	
education_unknown	
housing_no	
housing_yes	
job_admin.	
job_blue-collar	
job_entrepreneur	
job_housemaid	
job_management	
job_retired	
job_self-employed	
job_services	
job_student	

Summary Report:

Accompanying these Live Data reports in PowerBI will be a short, written report (word document) that outlines the following:

- Brief guide to navigating the Live Reports for the benefit of the Marketing Team
- Analysis of the features in the data and the outcome to recommend those features which ought to be used to satisfy the Operational objective
- The advantages and drawbacks to utilising both PBI and Python to meet the internal reporting requirements
- Challenges you encountered during processing of data and deployment of your reports
- Outline of your development process and technical documentation
- Future recommendations to Analysts working on this challenge

Supporting Resources

The project manager has provided you with a working example of a successful project that used Python to script a predictive model into PowerBI. In this example, it was a Decision Tree Classification model of mobile phone specifications. She advises that you implement the same structure into your own report. She has also requested an SME to walk you through an overview of how this was accomplished, including:

1. PowerBI Workbook – `PBI_DT_mobile_report.pbix`
2. Jupyter Notebook – `Python_DT_mobile_model.ipynb`
3. Guidance on scripting the python model in PowerBI - `DT_mobile_Guide.docx.`
4. Source data is `phone_classification.csv`

Assignment detail

Task 1 –Decision Tree Classification and Evaluation in Python

Load in the direct_marketing.csv dataset into Jupyter Notebook.

Clean and process its features.

The project manager advises that you learn and adapt the overall structure of `Python_DT_mobile_model.ipynb` to suit the needs of this project and its dataset.

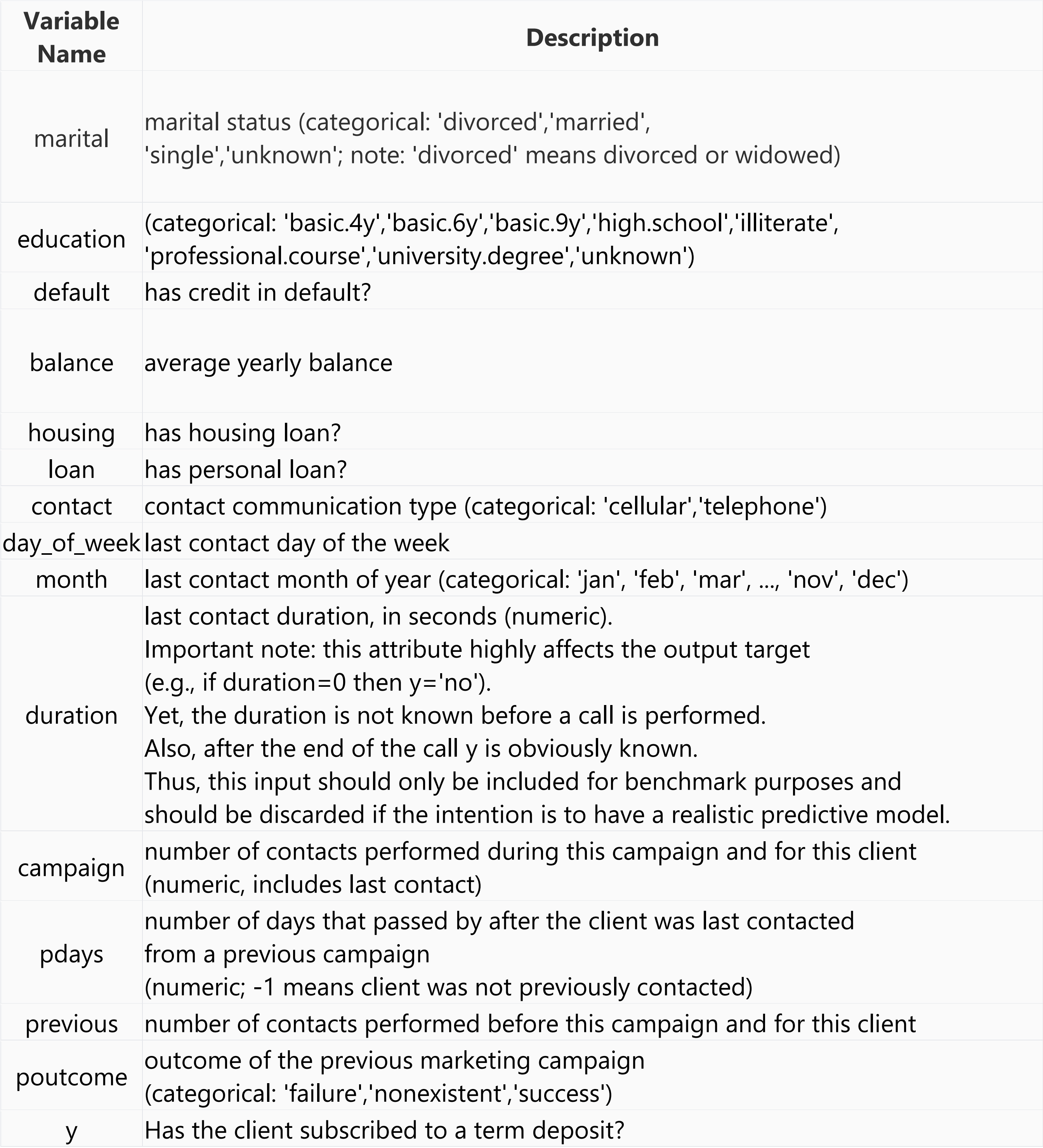
The goal is to build a functioning Classification model, visualise and evaluate it. The model should include its most informative features, a confusion matrix to evaluate the predictions and the model metrics (accuracy, recall and precision).

Key Requirement: You must One-Hot-Encode your categorical data. – encoding increases the compatibility with different modelling techniques and more accurate modelling is possible on high numbers of possible categorical values. One-Hot-Encoding ensures that during machine learning higher numbers are not treated as more important than lower numbers as the values represent categories in a non-ordinal sequence. This is only possible with a Python script.

REMINDER There are some features in the data that are only known after the campaign call is made. Most notably, Duration, which it the length of the phone call regarding the Term deposit product. Your model must be operational live so cannot rely on information that will be known after the marketing call.

Dataset Feature Glossary – *direct_marketing_dataset.csv*

Variable Name	Description
age	
job	type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid', 'management','retired','self-employed','services','student', 'technician','unemployed','unknown')



Variable Name	Description
marital	marital status (categorical: 'divorced','married', 'single','unknown'; note: 'divorced' means divorced or widowed)
education	(categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate', 'professional.course','university.degree','unknown')
default	has credit in default?
balance	average yearly balance
housing	has housing loan?
loan	has personal loan?
contact	contact communication type (categorical: 'cellular','telephone')
day_of_week	last contact day of the week
month	last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
duration	last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
campaign	number of contacts performed during this campaign and for this client (numeric, includes last contact)
pdays	number of days that passed by after the client was last contacted from a previous campaign (numeric; -1 means client was not previously contacted)
previous	number of contacts performed before this campaign and for this client
poutcome	outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
y	Has the client subscribed to a term deposit?

Task 2 – Embedding Python processing within PBI

This is where the bulk of your time will likely be spent. Much of the data objects (lists, models, arrays) you created in Jupyter must be formatted or combined into Data Frames for Power BI. This is the most compatible Python Data Object in PBI and facilitates all the processing necessary for the data and its modelling.

Refer to the `DT_Mobile_Guide.docx` that the project manager has provided to you as a key resource to help you complete this task.

Task 3 – Embedding Python Visualisations within PBI (Decision Tree Visual Report)

When you are satisfied with the data frames you have created in the Power Query editor and your decision tree model, save the transformations and move over to the PowerBI report view to begin creating your model Visualisation. You will visualise a Decision Tree for predictive modelling/ Classification, something that is only possible with Python or R within Power BI. The target Decision Tree visual output is shown on page 4 of this document.

Task 4 – Embedding Python model metrics within PBI (Evaluation and Features Visual Report)

Continue building tables and visualisations for the model evaluation in Power BI desktop. The final report must be well laid out, legible and contain appropriate visuals / tables. The target Evaluation and Features visual report is shown on page 4 of this document.

Task 5 – Summary Report

Summary report can be written as a document or assembled as a slide deck. Ensure that your summary report covers all required topics specified on page 4 of this document. Also ensure that your report is written in an audience appropriate manner – the audience in this scenario is your project manager, analyst colleagues and the marketing team.

Self-assessment

You will now self-assess your performance of the task against the below points schema:

- ✓ Task 1 - 4 points
- ✓ Task 2 - 5 points
- ✓ Task 3 – 2 points
- ✓ Task 4 – 2 points + 3 points (Advanced)
- ✓ Task 5 – 4 points

Potential total mark is 20.

Delivery

When you have completed this scenario, share your PowerBI book and summary report with your score to the trainer for review. If you have scored over 17/20, be prepared to showcase it briefly in class.