



Lloyds Banking Group

Data Analyst Incubation Program

Sprint 1 : Direct Marketing Campaign – Decision Tree Report





Contents

Introduction	3
Business need.....	3
Objectives.....	3
Communicating Results to stakeholders.....	4
Decision Tree Visual Report:	4
Evaluation and Features Visual Report:.....	4
Summary Report:	5
Supporting Resources.....	5
Assignment detail	6
Task 1 –Decision Tree Classification and Evaluation in Python.....	6
Task 2 – Embedding Python processing within POWER BI.....	7
Task 3 – Embedding Python Visualisations within POWER BI (Decision Tree Visual Report).....	8
Task 4 – Embedding Python model metrics within POWER BI (Evaluation and Features Visual Report)	8
Task 5 – Summary Report.....	8
Self-assessment	9
Delivery.....	9



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 called **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 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

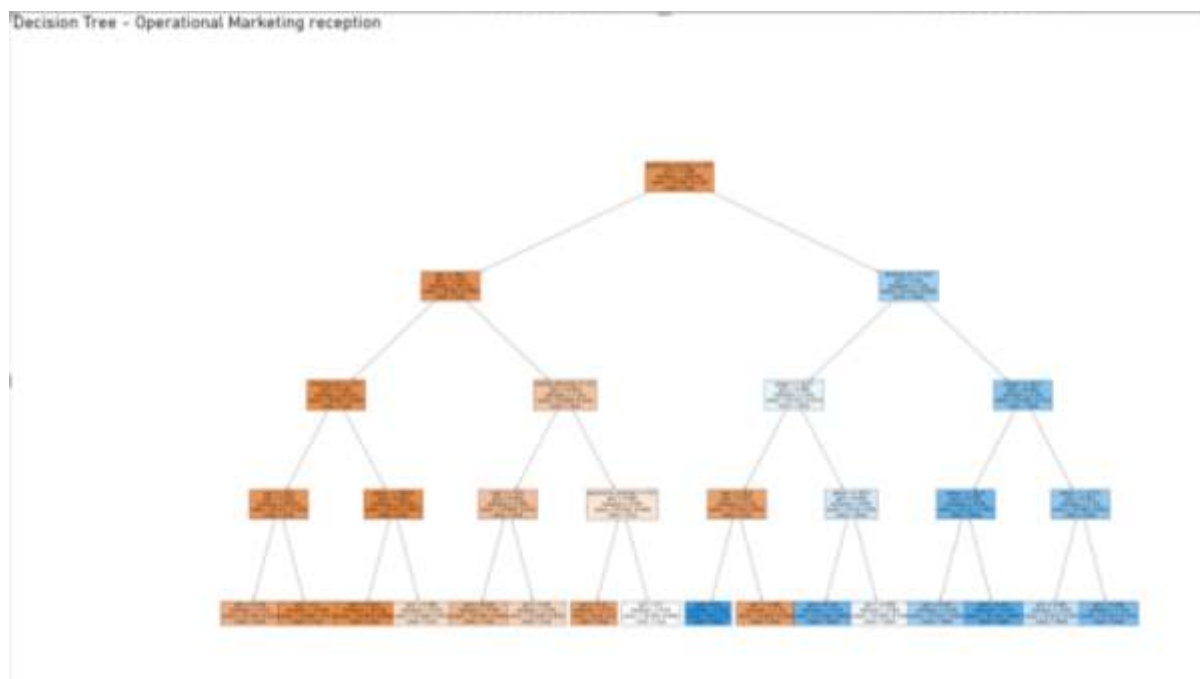
To meet the stakeholder requirements, it is required that you produce 3 reports to share your results. These are described in more detail below.

1- 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 report view, select fill page or fit width. In Python, you would use the matplotlib library to alter **fig(width, length)** and font size to achieve the desired legibility.



2- Evaluation and Features Visual Report:

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

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



3- Summary Report:

Accompanying these Live Data reports in Power BI 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 POWER BI 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 Power BI. 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. Power BI Workbook – **POWER BI_DT_mobile_report.Power BIx**
2. Guidance on scripting the Python model inside Power BI – **DT_mobile_Guide.docx.**
3. 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 any previous machine learning models you have developed 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 is 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	age in complete years
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'). Reminder : the duration is not known before a call is performed and at that time the outcome is also known. Therefore this feature should be included for benchmark purposes only and should be discarded if the intention is to have a realistic predictive (operational) 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 Power BI

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. Data frames are the most compatible Python Data Object in Power BI for reports.

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

Task 3 – Embedding Python Visualisations within Power BI (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, move over to the Power BI 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 Power BI (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

Your 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 Power BI report 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.

