In this Lab, you’ll practice your Python and machine learning skills as you progress through the Data Analytics Workflow to dissect a business problem. The Lab consists of two parts, and in each part you’ll use Python to explore and model data.

# Scenario and Timing

You are a data analyst working for a Portuguese bank. The bank has recently run a campaign to attempt to sell a new product to customers over the phone. The product is a [term deposit](https://www.investopedia.com/terms/t/termdeposit.asp#:~:text=A%20term%20deposit%20is%20a,levels%20of%20required%20minimum%20deposits.), where customers deposit their money for a fixed period of time for a fixed interest rate. They want to analyze the results of this campaign to understand what makes customers likely to buy this particular product. Your job as the data analyst is to explore the data and use machine learning to build a model to predict whether a customer is likely to buy.

The data dictionary can be found in the Appendix of this document.

You will have **a day** of class to complete this Lab, and you need to complete **both** parts.

# Part 1: Exploration (time: half day)

In this section, follow along with Section 1 of the Jupyter notebook to clean and explore the data. Your focus will be on:

* Investigating descriptive statistics about your columns.
* Identifying missing values and dealing with them accordingly.
* Exploring the relationship between your potential features and the target, e.g. answering questions like, “How did the percentage of people who bought the product vary with the age of customers?” and “Are older or younger customers more likely to buy?” Your answers to these questions will help you choose variables for the machine learning model you will build in Part 2.

# Part 2: Modeling (time: half day)

Using the results from Part 1, you will follow along with Section 2 of the Jupyter notebook and train a machine learning model to predict whether a customer will buy the banking product. You will then analyze the performance of your model, and decide on ways to improve its performance, for example by adding or removing features.

# Deliverables

1. A completed Jupyter notebook (\*.ipynb) with the required Python code filled in, exploring the data and training **two** machine learning models.
2. A completed 2 page Word report describing the results of the analysis for each part. Your report should include the following sections (the template is [here](https://docs.google.com/document/d/1kmWTQPvaV6dY7pqxIi7E6jSc6egfopDd71RRW_c100w)):
3. **Part 1 - Exploration** which explains the findings in section 1 of the notebook. In particular, it should detail which variables you chose for your models and why.
4. **Part 2 - Modeling** which explains the results of your machine learning models and a comparison of the two models you built: how their performances compare, and which is the best model. Discuss the performance in a way your stakeholders could understand, e.g. by discussing what false positives and false negatives mean for your problem.

# Process

Your finished deliverable should include evidence of:

## Cleaning the Data

* + The report should mention any missing data you encountered and what you decided to do with it.

## Exploring the Data

* + Which factors affect the likelihood of a customer buying the marketed product? Exploring this question for each of your potential features will inform your decision when building your machine learning model.

## Modeling the Data

* + How well did your model perform? Include relevant success metrics as well as a discussion of what they mean for the business context, (e.g. how do you know if a particular accuracy score is “good”?).
  + Based on the results of the first model, what did you change before building the second one? What were the consequences? Which model performed better?

# Rubric

The purpose of this assignment is to assess your Python proficiency. Instructors will evaluate student skill based on the following rubric:

|  | **Incomplete** | **Doesn’t Meet Expectations** | **Meets Expectations** |
| --- | --- | --- | --- |
| **Jupyter Notebook** | Notebook not submitted | Notebook submitted but not completed | Notebook completed and submitted |
| **Report**  ***Word / Google Doc*** | Document not submitted | Document submitted but does not answer both parts | Document submitted and answers both parts |
| **Data Cleaning** | Data not cleaned/modified in any way | Data is partially cleaned (e.g. missing values dropped without justification) | Data has been cleaned and decisions have been justified in the report |
| **Analysis** | No analysis of the data has been performed | Some analysis has been done but is incomplete | Analysis has been performed, features chosen for the machine learning section, and results communicated in the report |
| **Modeling** | Modeling section not complete | Modeling section incomplete (e.g. no discussion of results, no second model built for comparison) | Modeling has been completed, results have been discussed and two models have been compared against each other |

You must receive a score of at least **Meets Expectations** in all categories to pass this Lab.

# Good luck and have fun!

# 

# Appendix: Data dictionary

| **Column** | **Description** |
| --- | --- |
| **age** | Age of the customer in years |
| **job** | Customer’s job type |
| **marital** | Customer’s marital status |
| **education** | Customer’s education level |
| **default** | Did the customer have credit in default at time of contact? |
| **housing** | Did the customer have a housing loan at time of contact? |
| **loan** | Did the customer have a personal loan at time of contact? |
| **contact** | Contact communication type |
| **month** | Month of the year of latest contact |
| **day\_of\_week** | Day of the week of latest contact |
| **campaign** | Number of contacts performed during this campaign and for this customer |
| **pdays** | Number of days that passed by after the customer was last contacted from a previous campaign (999 means client was not previously contacted) |
| **previous** | Number of contacts performed before this campaign and for this customer |
| **poutcome** | Outcome of the previous marketing campaign |
| **emp.var.rate** | Quarterly economic indicator (employment variation rate) at time of contact |
| **cons.price.idx** | Quarterly economic indicator (consumer price index) at time of contact |
| **cons.conf.idx** | Quarterly economic indicator (consumer confidence index) at time of contact |
| **euribor3m** | Euribor 3-month rate. European central bank interest rate, can be used as a proxy for the state of the economy at time of customer contact |
| **nr.employed** | Quarterly economic indicator (number of employees) at time of contact |
| **y** | Target variable - has the customer subscribed to a term deposit? |