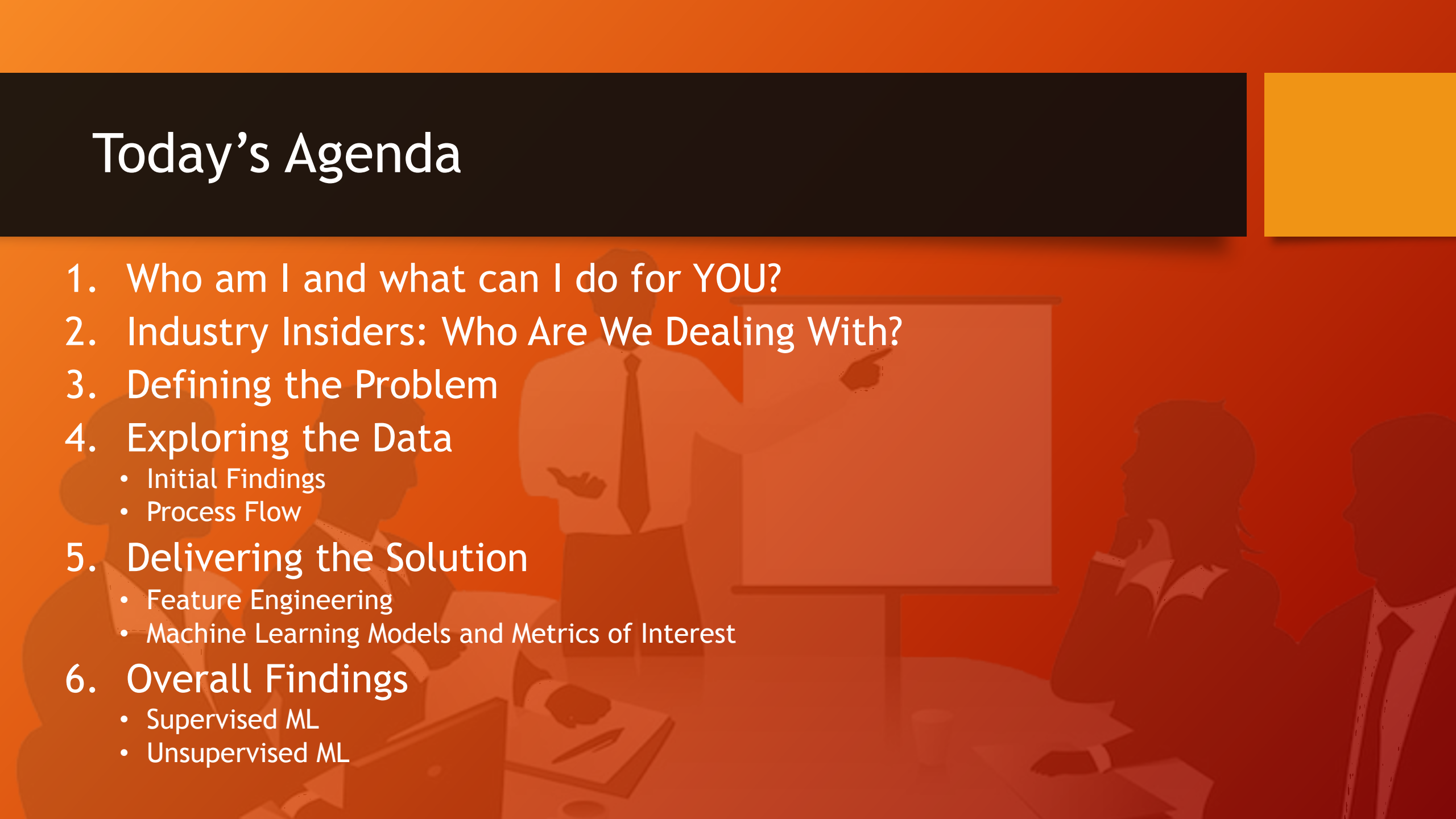
Telemarketing Success

Institute of Data Capstone Project

Anna-Maria Schreiner, Aspiring Data Scientist

Today's Agenda

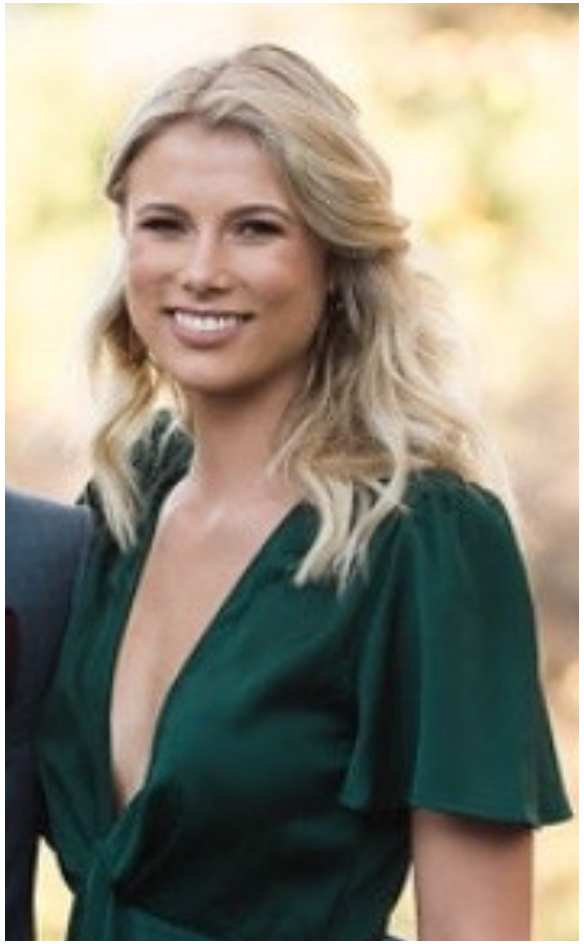
1. Who am I and what can I do for YOU?
2. Industry Insiders: Who Are We Dealing With?
3. Defining the Problem
4. Exploring the Data
   - Initial Findings
   - Process Flow
5. Delivering the Solution
   - Feature Engineering
   - Machine Learning Models and Metrics of Interest
6. Overall Findings
   - Supervised ML
   - Unsupervised ML



What Can I do For You?

- **Educational Background:**
  - Bachelor of International Business Management at University of Queensland (Australia) and Universität St Gallen (Switzerland)
  - RMIT/ Udacity Business Analytics Nanodegree
  - Institute of Data Graduate Certificate of Data Science and Artificial Intelligence

- **Data Science Learning and Experience**
  - Immersive training alongside real-time application to continuously build a portfolio of experience.



Industry Insiders: Who Are We Dealing With?

- Client:
  - Central bank of the Portuguese Republic (Banco de Portugal)
- Problem area:
  - Problem -> Declining revenue
  - Strategized Solution -> deploying direct telemarketing campaigns to promote long-term deposits to increase
- How can I help?
  - Modeling the success of the marketing campaigns will be indispensable knowledge gained and add significant value:
    - ✓ Success rate of the campaign (present and future)
    - ✓ which clients are most financially viable to target
    - ✓ Optimize marketing strategies and improve effectiveness

Time to Solve Some Problems!

Business Question

- *Which existing customers would have a higher probability of responding positively to a long-term deposit marketing campaign?*

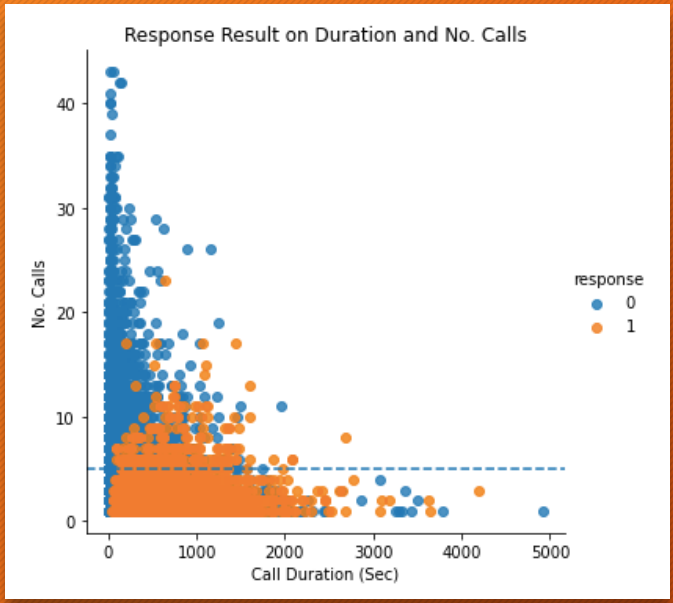
Data Question

- *Can a Machine Learning model predict which clients are most-likely to successfully respond to telemarketing calls aimed at selling long-term deposits?*

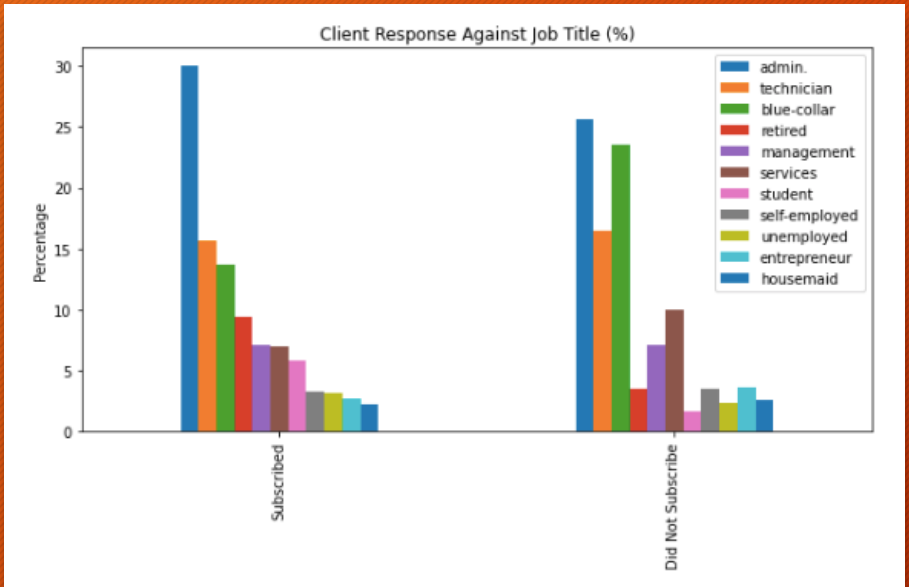
Desired Output

- *Enable management to develop a more granular understanding of their customer base and predict the customers response to marketing strategies.*

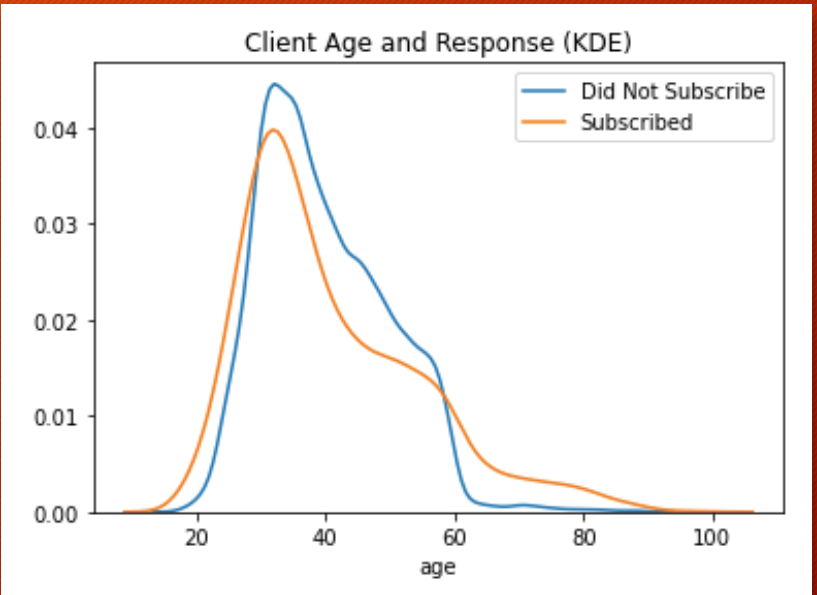
1.



2.



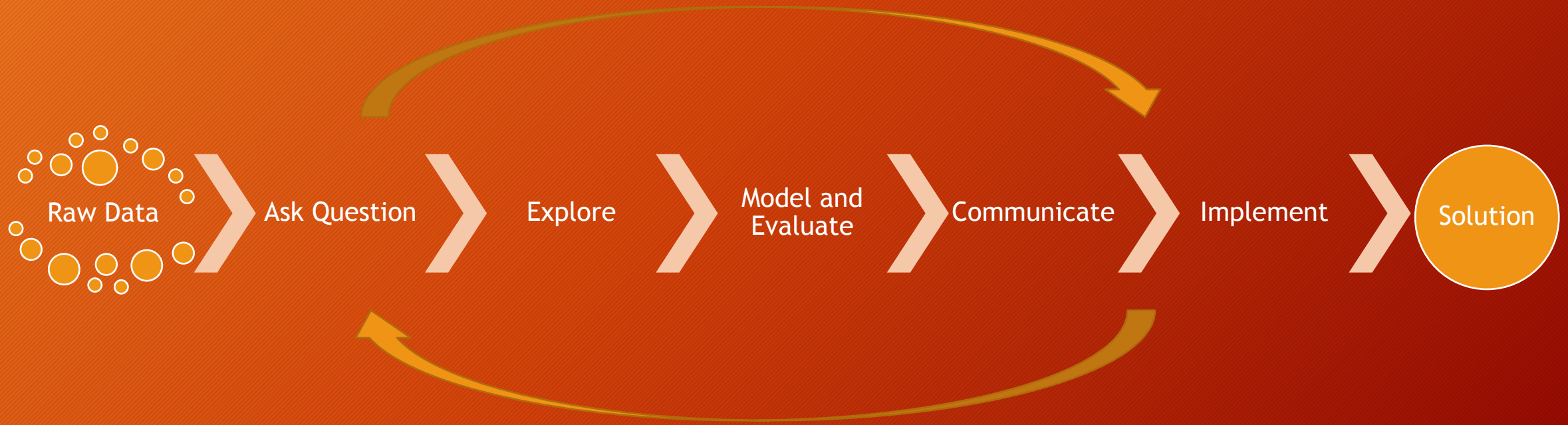
3.



Exploring the Data - Initial Findings

Exploring the Data
- Data Science Process Flow

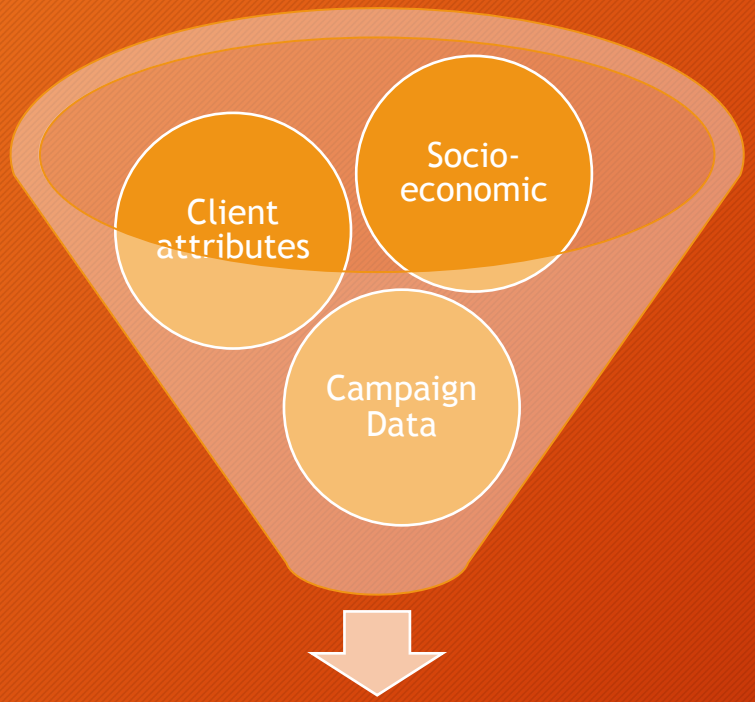
To be iterated in line with the business and data questions, via stakeholders



***Potential misunderstanding, should this be: data cleaning, removing nulls, eliminating outliers, analyzing correlations?

Delivering the Solution - Feature Engineering

Inputs: Predictor Variables



Output: Client Response (Binary Classification)

Preparing the Feature Vector:

1. To maintain a strong model, most features were used - only those that deemed inaccurate or incomplete were removed.
2. Categorical attributes had to be converted numerical by mapping binary variables or deploying dummy variables
3. Normalized scaler to standardize dummy variables with original numerical data
4. SMOTE: Synthetic sampling technique for class imbalance

Delivering the Solution - ML Models and Metrics

Binary Classification Modelling (Supervised):

1. Logistic regression
   - With and without synthetic resampling (SMOTE)
2. Support Vector Machine

Comparison Metrics:

- Receiver Operating Characteristic (ROC Curve): Specificity vs Sensitivity
- Accuracy Score: fraction of predictions the model got right
- Recall: proportion of actual positives correctly identified (sensitivity of the classifier model)
  - TP/(TP + FN)
- Precision: the proportion of correct positive identifications
  - TP/(TP + FP)

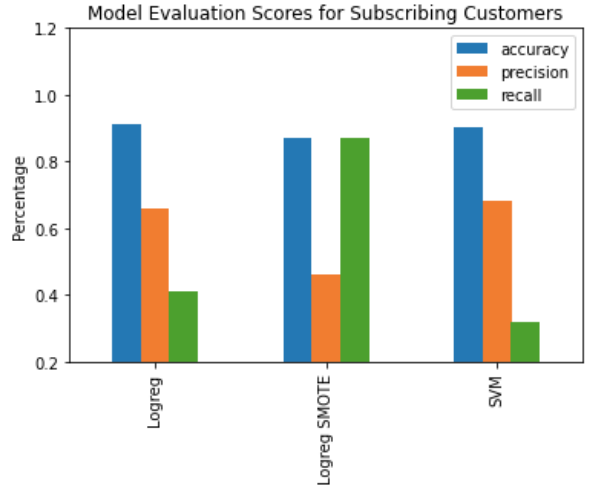
Clustering Challenge! (Unsupervised):

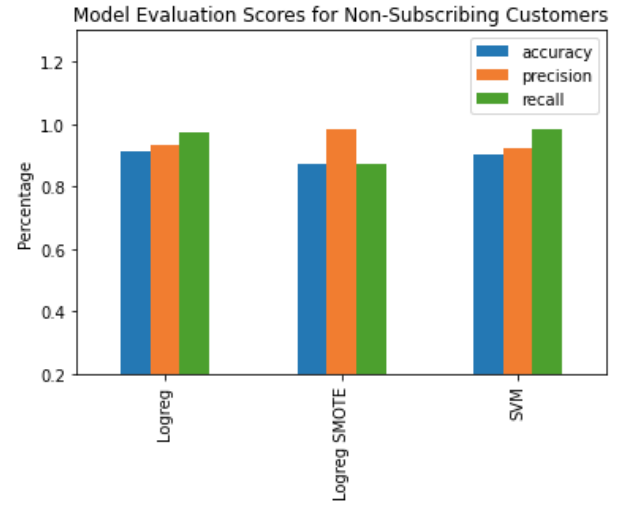
1. K-means
   - An iterative algorithm that attempts to partition the data into distinct subgroups
   - Will one cluster of client attributes be likely to respond more positively to the telemarketing campaign than another?

(1) Findings: Supervised ML

- Summary: Model Evaluation
  - Accuracy Score maintained consistency but deemed potentially unreliable on an unbalanced classifier dataset. However, accuracy of below 90% implies less risk of overfitting the data (increasing generalizability).
  - Logistic regression was comparatively insensitive in its detection of classes UNTIL the data was synthetically resampled (SMOTE) (From 41% recall to 87%!)
  - SVM had stronger precision but significantly low recall in detecting positive responses (subscribing clients)
- Next Steps: Stakeholder Value
  - Using the most diligent model, LogReg SMOTE, Banco de Portugal can predict the responses of their clients in the context of future telemarketing campaigns and strategize accordingly.

	accuracy	precision	recall
Logreg	0.91	0.66	0.41
Logreg SMOTE	0.87	0.46	0.87
SVM	0.90	0.68	0.32





(2) Findings: Unsupervised ML

- Summary: K-Means Customer Profiling (when optimal k=2)

CLUSTER 1	CLUSTER 2
• Less likely to subscribe • Age: 40+ • Education: Primary/ base education • Marital: Married • Job: Blue-collar roles	• More likely to subscribe • Age: Less than 40 • Education: Tertiary • Marital: Highest representation of singles • Job: administration and technician roles

- Next Steps: Stakeholder Value
  - Optimized marketing strategies = more efficient targeted approach
  - Understanding customer needs = leads to more effective campaigns, smarter product design and greater overall customer satisfaction.

✓ Successfully turned data into information, and information into insight!