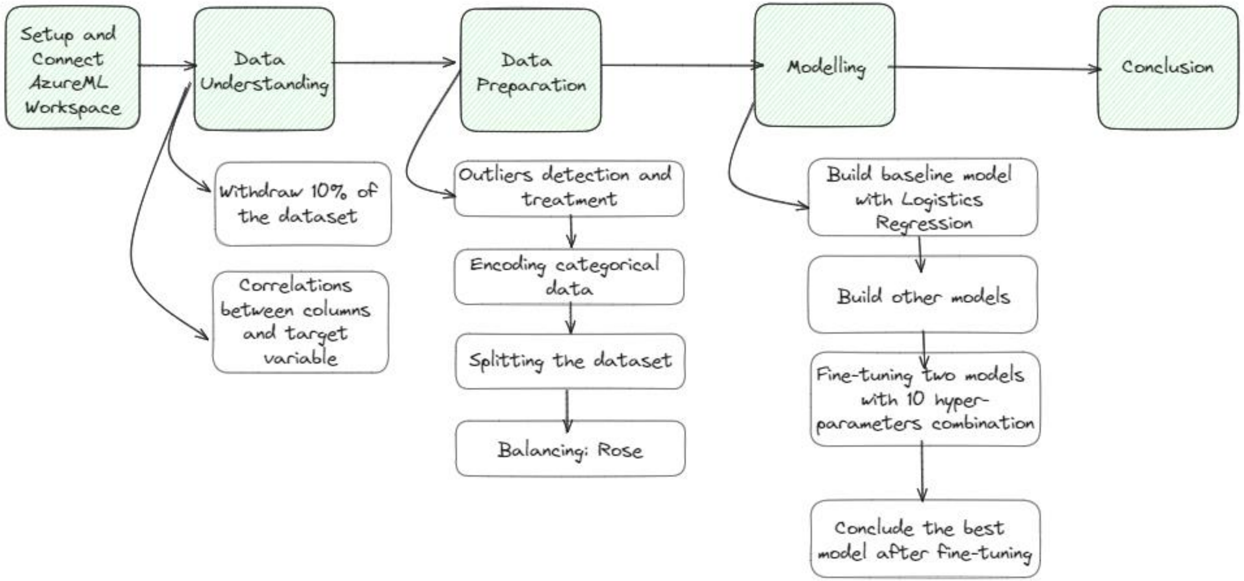
Methodology and Workflow

- The methodology applied for this dataset is based on the CRISP-DM Model.
- The basic workflow is illustrated as followed:


Setup and Connect AzureML Workspace
Data Understanding
Data Preparation
Modelling
Conclusion
Withdraw 10% of the dataset
Correlations between columns and target variable
Outliers detection and treatment
Encoding categorical data
Splitting the dataset
Balancing: Rose
Build baseline model with Logistics Regression
Build other models
Fine-tuning two models with 10 hyper-parameters combination
Conclude the best model after fine-tuning

About the Dataset and Insights

- The dataset details phone-based marketing campaigns by a Portuguese bank promoting term deposits, involving multiple client contacts to determine subscriptions

THE DATASET

- Given that the full dataset is extensive, comprising 17 attributes and encompassing 45211 rows , a smaller dataset is extracted by the 10% of the dataset to serve as a sample for analysis and model building. The sample dataset is consisted of:

4,452 rows	17 columns
87.96% no	12.04% yes

INSIGHTS

Most clients fall into the 40 age group, yet their success rate is low. Only 4% are 60 or older, while 11% are under 30; these age brackets show higher success rates

The top contacted clients are in job types 'blue-collar,' 'management,' and 'technician,' but blue-collar clients have a low success rate. Students have the highest success rate

The majority of contacted clients hold tertiary or secondary education levels, with both groups exhibiting high success rates

May sees the most client contacts (approximately 1/3 of the total), but with a success rate of only 6.7%. March boasts the highest success rate despite being one of the least contacted months

The majority of clients are contacted via cell phone.

Most successful clients do not have a loan, with only a small percentage of contacted clients who already have one achieving success

Primary Evaluation

- In total, 10 models are built to predict the dataset. Here are the final results in terms of Accuracy and AUC indicators.

baseline

	Accuracy	AUC
Logistic Regression	0.8058	0.7726
SVM	0.5364	0.5389
MLPClassifier	0.8462	0.7028
AdaBoostClassifier	0.8350	0.7448
KNeighborsClassifier	0.6498	0.595
RandomForestClassifier	0.8754	0.6225
DecisionTreeClassifier	0.82828	0.7451
GaussianProcessClassifier	0.8103	0.5613
HistGradientBoostingClassifier	0.8574	0.7455
GradientBoostingClassifier	0.8631	0.6922
GaussianNB	0.7935	0.6526

- As can be seen, the baseline model performs well compared to the other models.
- Most of the models achieve notably high accuracy scores.
- For experimental purposes, among the models, Random Forest and Decision Tree perform quite well in terms of accuracy and AUC score, so I will use these two models along with Logistic Regression for fine-tuning.

Final Evaluation

- After fine-tuning, here are the new table results:

best model

	Accuracy	AUC	Best hyperparameters
Logistic Regression	0.8171	0.7911	(0.1, 'liblinear')
DecisionTreeClassifier	0.8507	0.6852	(2, ‘gini’)
RandomForestClassifier	0.8866	0.6894	(100, ‘gini’, ‘auto’)

- All of the models yield higher results after fine-tuning.
- Although both Random Forest and Decision Tree show relatively high accuracy scores, their AUC scores are quite low, at 0.66 and 0.68 respectively.
- Therefore, Logistic Regression with the combination of hyperparameters (0.1, 'liblinear') will be chosen as the best model, as it can achieve a balance between an accuracy of 0.81 and an AUC score of 0.79.

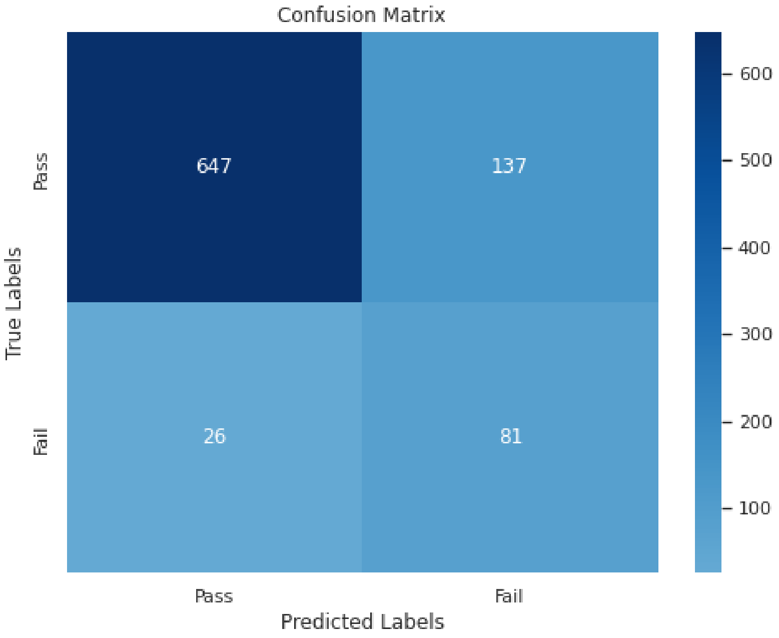
About the best model:

Logistic Regression (0.1, 'liblinear')

- The model seems to perform reasonably well, with decent accuracy (81.71%) and AUC score (79.11%).
- However, the measure of the proportion of true positive predictions out of all positive predictions is relatively low, indicating that there may be a higher rate of false positives.

Next steps:

- Further tuning or feature engineering might be necessary to improve the model's performance, especially if reducing false positives is crucial in the context of the application.
- Perform the model with the full dataset.
- Examine and fine-tune the dataset with other classification models.



Reference

- https://github.com/nickr007/Bank-Marketing/blob/master/Bank%20Marketing.ipynbhttps://chat.openai.com/share/86bd6096-7b2d-4fed-b1d6-397c56dd2eed
- https://chat.openai.com/share/007b67b2-7374-46c4-9fa9-09d9ca14d957
- https://chat.openai.com/share/3df566a6-1a1e-4180-89df-f6c4bf3b5881