**Purpose**

This report provides a high-level summary of the model that predicts if a client will subscribe to a term deposit.

**Analysis Summary**

The model was developed using Logistic Regression because the intention is to predict one of two possible outcomes and because the target variable is categorical. The dataset titled bank-additional.csv was used because it is the most complete and most recent of the four datasets provided. This data is however imbalanced with 36548 records of clients who did not subscribe and 4640 records of clients who subscribed.

After exploratory data analysis, one-hot encoding and feature selection were used to identify the variables that determine a client’s likelihood to subscribe to a term deposit.

The initial model built on the original dataset had an accuracy of 90% but with a precision score of 66% and a recall score of 20%. To improve the model, under-sampling, oversampling, combined under-sampling and over-sampling, and class weight techniques were used. The model built after adjusting class weights had the highest accuracy and recall amongst the other techniques, therefore it was chosen.

The chosen model had an accuracy score of 79%, a precision score of 30%, and a recall score of 66%. This means the model accurately captures 66% of the clients who will subscribe to the term deposit. Even though precision is low, this model had the highest recall percentage, hence why it was chosen.

**Insights**

The features that are most impactful in determining if a client will subscribe to a term deposit are:

* The number of contacts performed during this campaign
* The number of contacts performed before this campaign
* The employment variation rate of the client
* The Euribor 3-month rate
* The client’s job and education
* Whether the client has credit in default
* The contact communication type,
* The month and day of the week the client is contacted
* The outcome of the previous marketing campaign

The recommendations for the marketing team are:

* Focus on clients with a history of successful campaigns
* Prioritize outreach during high-impact months (March, December, August, and July)
* Target students and retired clients
* Improve Messaging for Clients with No Previous Contact Outcome

The limitations for the model are:

* The dataset used is imbalanced.
* The model is heavily skewed toward predicting no
* Logistic regression is a linear model and assumes a linear relationship between predictors and the log-odds of the outcome.