An Analysis on Marketing Strategy for N/LAB Platinum Deposit by N/LAB Enterprises

INTRODUCTION

N/LAB Enterprises is seeking to expand its operations into the banking industry and has developed a strategy to generate capital. This strategy involves marketing a financial product called the **"N/LAB Platinum Deposit"**, which offers a fixed-term interest rate to individuals who make a large deposit that cannot be withdrawn for a year.

This report aims to summarize how N/LAB Enterprises should choose a customer to contact and onboard them to their new product. Using historical data of a similar product that N/LAB Enterprises has taken over, we can achieve this by performing data analytics and predicting customers who are willing to opt for the product.

As requested by the CEO, the final model should be able to accurately label an individual who is not interested, because the cost of fruitless calls is more. The CEO is okay even if the number of individuals contacted are less.

SUMMARIZATION

Firstly, to understand historical data and explore how they provide insights on a potential customer, we will look at how different parameters in the dataset are linked to an individual agreeing to invest or not.

The data has 4000 rows with 8 categorial features and 7 numerical features. The data is skewed towards individuals not interested, i.e., there are more rows with output feature value as 'no'

Visualizing how different parameters are linked to the target feature, can help us initially check if the data provided can help us predict future outcomes and how important these parameters are. So, considering the following plots.

The most evident feature that distinguishes the output feature is the last contact **duration**. As seen below, the box plot shows that if the last call lasted more than 400 seconds, then the individual is more likely to participate in the fixed-term deposit plan.

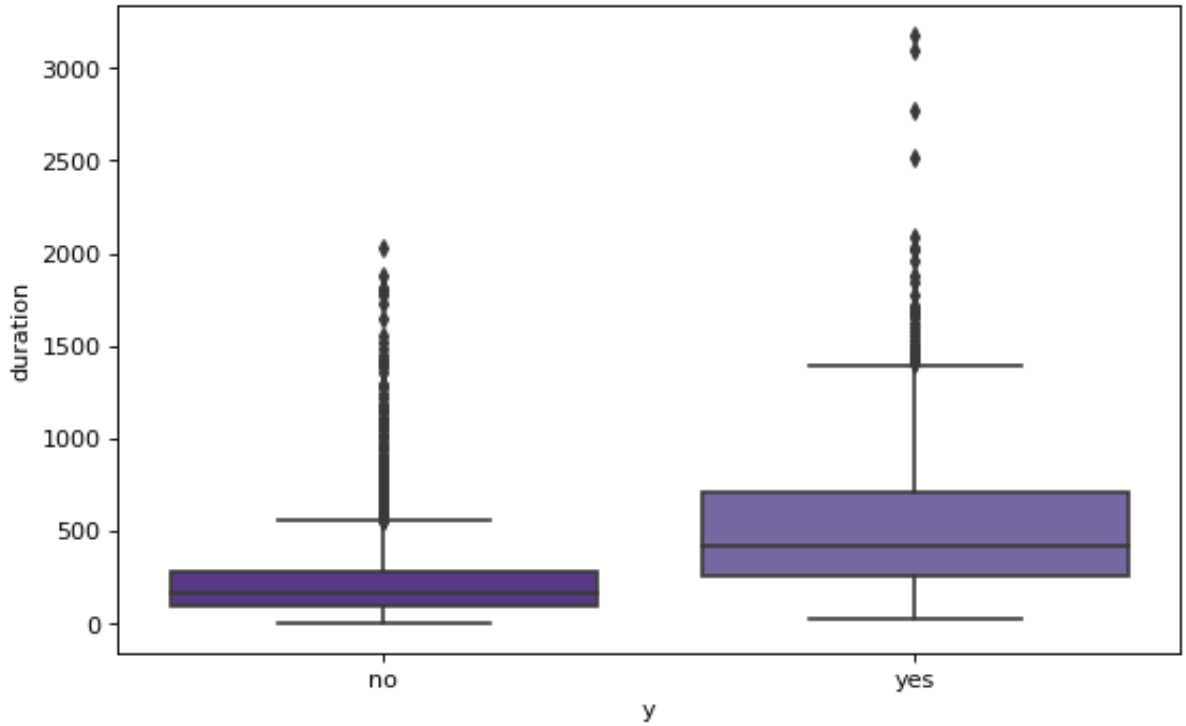

Figure 1: Box plot of duration vs output feature

The number of days (**pdays**) that have passed after the individual was contacted in a previous campaign is a good indicator as well as we can see in the below chart. If the outcome (**poutcome**) of such calls was a success, the individual is more likely to say yes and if it was a failure, one is more likely to say no.

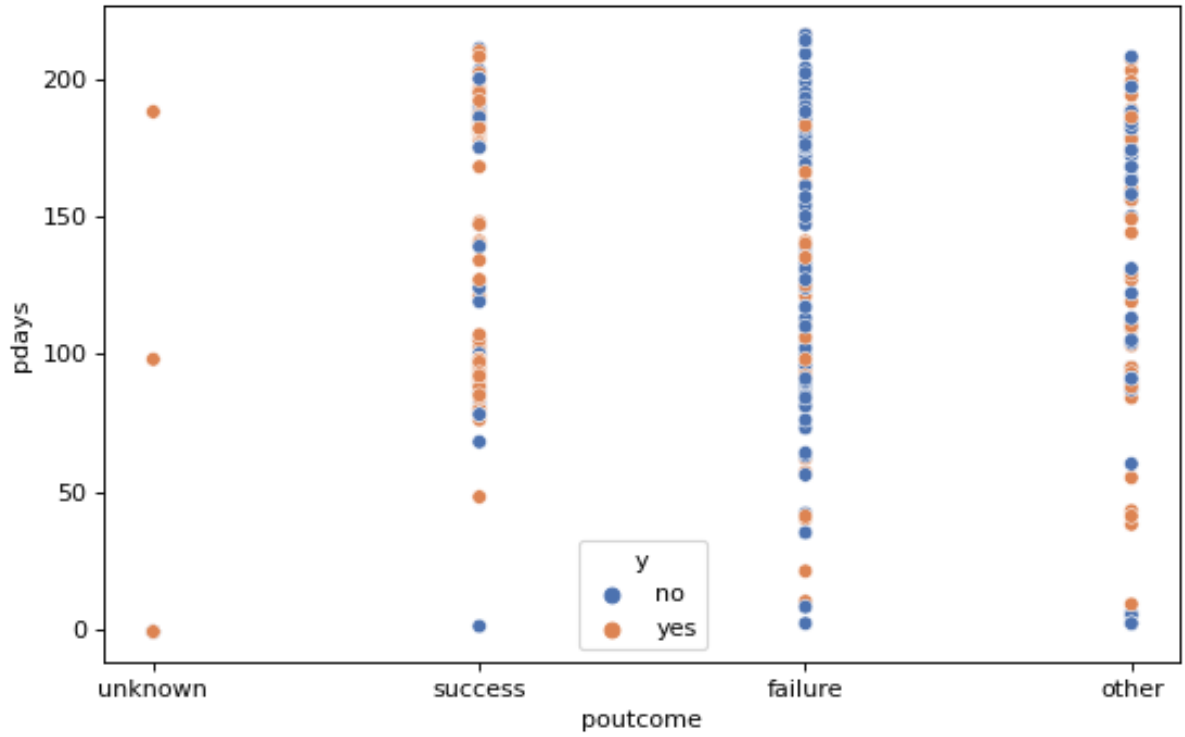

Figure 2: Scatter plot of pdays vs poutcome with hue as output

If the individual has taken a **housing** loan, then the individual again wouldn't want to invest in a long-term deposit as shown by the violin plot below. The distribution of 'no' is more towards having a housing loan and the distribution of 'yes' is more towards not having a housing loan.

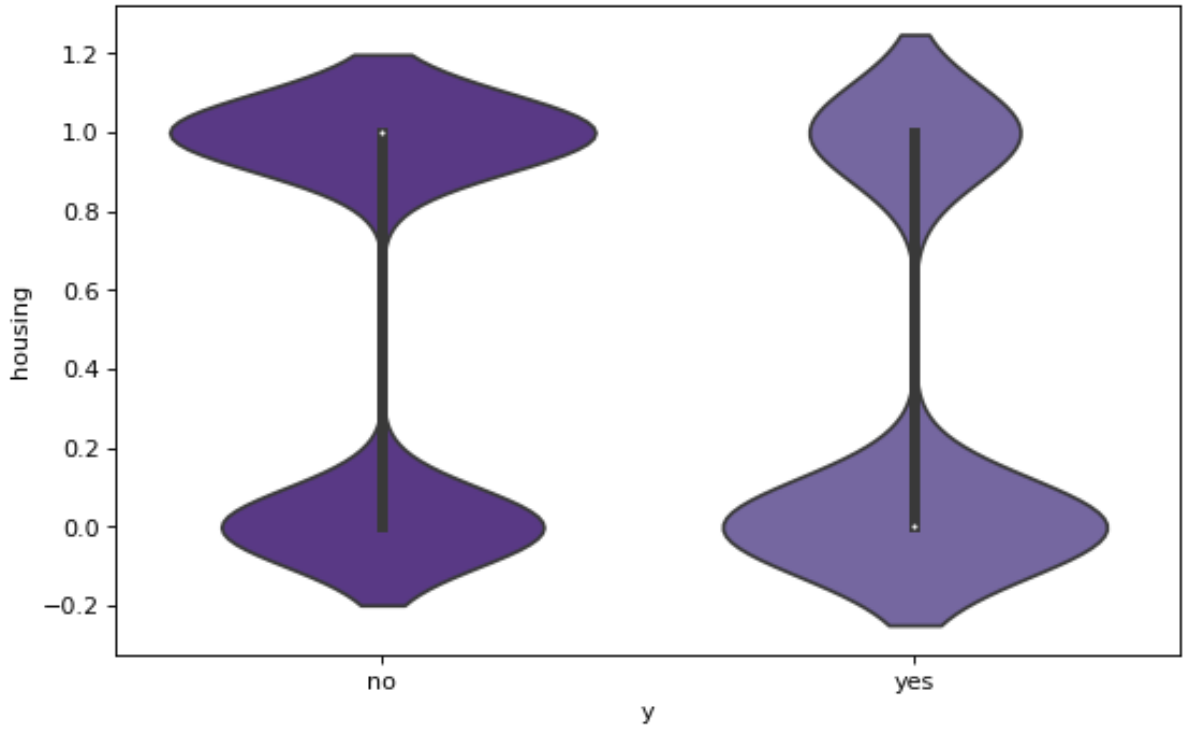

Figure 3: Violin plot of housing vs y (0 is not having a housing loan)

Now, let us also check how the input features are related to each other by drawing a correlation matrix as shown below.

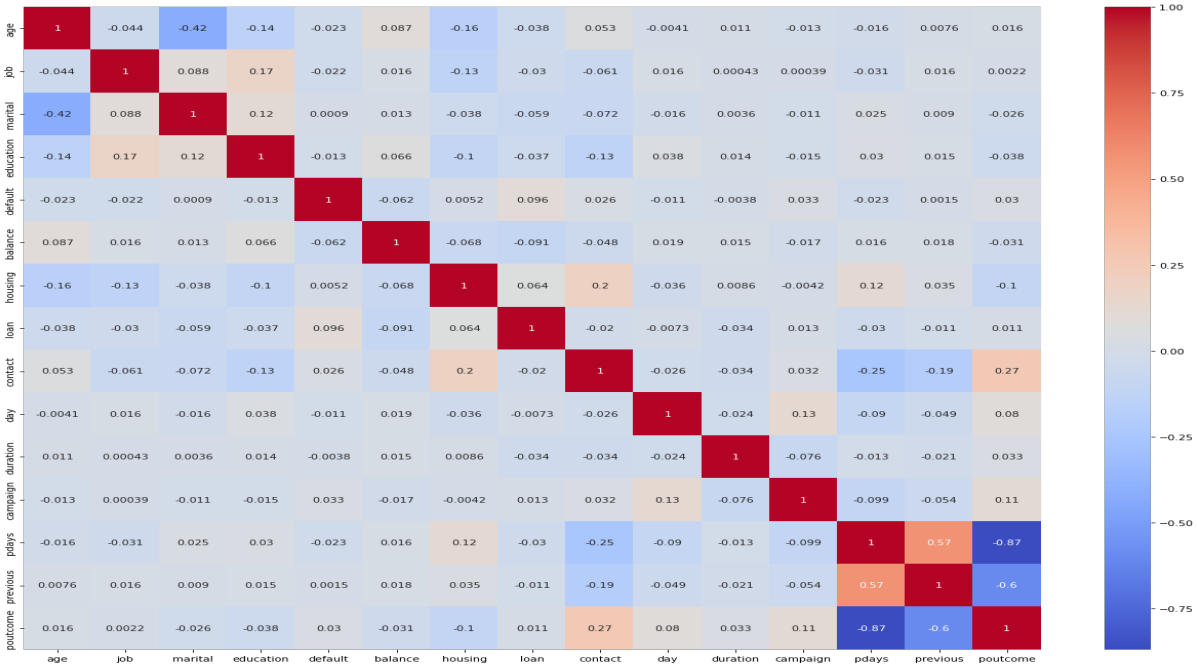

Figure 4: Correlation matrix

As we can see, pdays, previous and poutcome are highly correlated but that is mainly because there are a lot of individuals in the dataset who have never been contacted in a previous campaign. Hence directly removing a correlated column is not a good choice in this scenario.

Now that we have a good idea about how the historical data is structured, we can confirm the most important features by fitting the data into a decision tree model.

EXPLORATION

Initially, a decision tree model can be used as a classifier, because it provides a good visualization for us to check which features are providing the most information. And how multiple features together help us identify potential customers easily.

Cleaning the data: Before plugging it into the decision tree, firstly, categorical variables must be encoded, and outliers must be removed. To encode the data, an **ordinal encoder** can be used which uses numbers to differentiate between categories. The features are not converted into dummy variables here because apart from the job feature, all of them have a maximum of 4 categories.

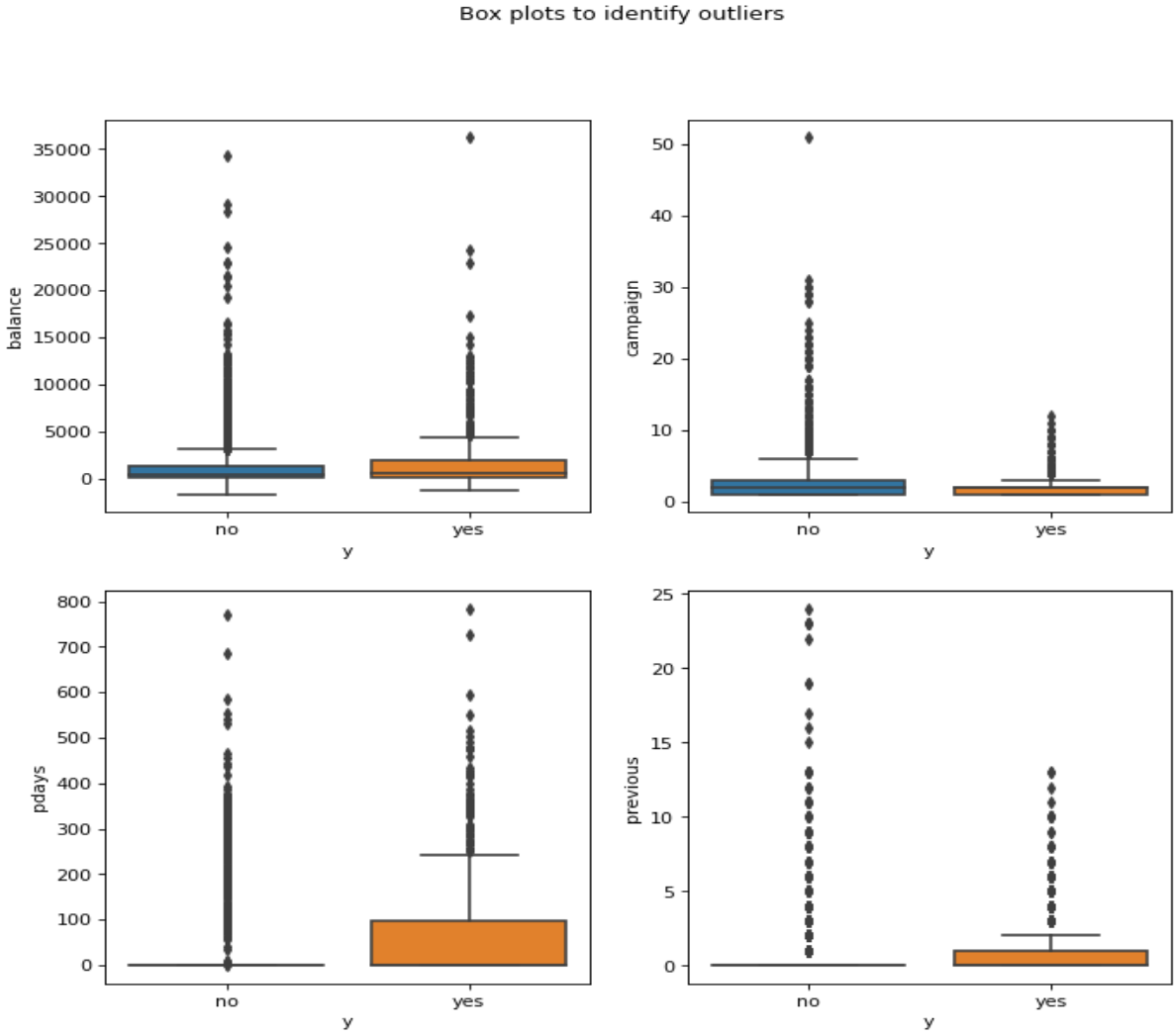

Figure 5: Box plots to identify outliers

Among these features outliers after a considerable gap are removed to make the data less scattered. Removing the outliers above 4 features results in our historical data being reduced to 3763 rows.

Fitting this into a decision tree of max-depth 4, we can visualize the features which have a lot of influence in deciding if the individual chooses the fixed-term deposit plan.

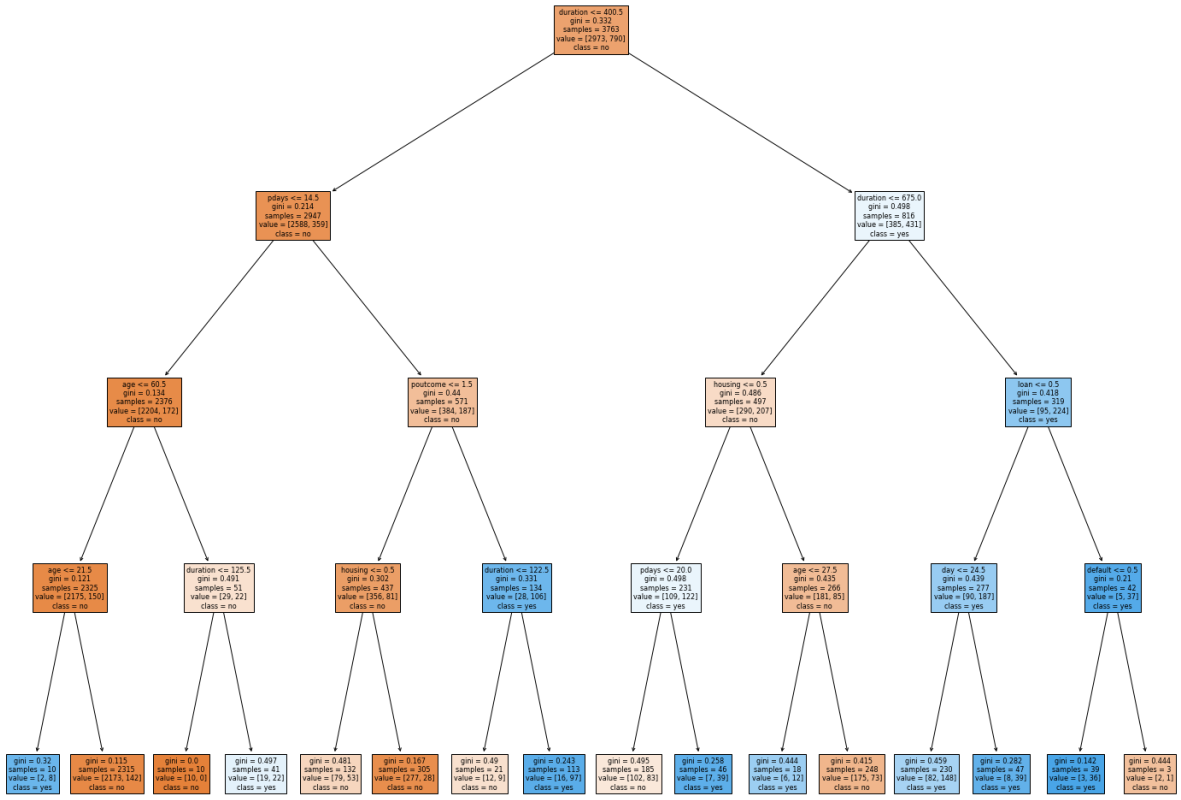

Figure 6: Decision tree with max-depth 4

As analysed initially **duration** is the most informative feature, hence the decision tree has chosen it as the root node. Also **pdays** and **poutcome** together help classify individuals accurately. If a person has an **housing loan** or a **personal loan**, they are less likely to say yes to the new plan. Personal loan is a new feature which can be identified from the decision tree as important which was missed in simple visualizations. Also **age** plays an important role in determining the output feature.

Although predominantly duration is the most informative feature, we will have to remove this feature before training a machine learning model since the duration of a call cannot be known before contacting a potential customer.

MODEL SELECTION

To define the final model which predicts the chances of an individual choosing the plan, different classifying models can be experimented. The model that best meets the requirement will be chosen finally.

Defining the target class: Since the CEO wants to avoid fruitless calls which are more costly, the target class will be 'no'.

The following 3 models will be chosen for experimenting with reasons respectively:

- **Logistic Regression:** The logistic regression is a simple and efficient classification model. Very easy to set up and analyze for data with binomial classification. This model will also not overfit the data.
- **Random Forest:** Random Forest model is also resistant to overfitting the data since it uses multiple decision trees to randomly select each split. It can handle high dimensional data better than a logistic regression model. It is also robust to noise in the data.
- **K-Nearest Neighbors:** kNN model simply memorizes the training data and makes predictions based on the number of nearest neighbors defined. This model handles multi dimensions easily. Although it is slow to memorize huge sets of data, it classifies new data points in an efficient way.

PERFORMANCE OF EACH MODEL

Logistic regression: Fitting the cleaned and encoded data into a logistic regression provides an accuracy score of **79%** with the below confusion matrix. max_iter parameter is provided a value of 2000 for this model, to avoid the model getting stuck at a local maximum.

No		**2934**	**39**
Yes		**719**	**71**
Test	Predicted	No	Yes

The recall is given by the following formula:

$$Recall = \frac{True\ positive}{True\ positive + False\ negative}$$

A positive class in this case is 'no' since that is the target class. Hence, from this formula the recall is **98.68%.**
The number of correctly classifying as 'yes' is **71.**

Random Forest: Fitting the cleaned and encoded data into a random forest provides an accuracy score of **83%** with the below confusion matrix. Max-depth is provided to the model as 5 so that the model does not overfit the data.

No		**2935**	**38**
Yes		**597**	**193**
Test	Predicted	No	Yes

The recall is **98.72%.**
The number of correctly classifying as 'yes' is **193.**

K-Nearest Neighbors: Fitting the cleaned and encoded data into a kNN provides an accuracy score of **83%** with the below confusion matrix. The number of nearest neighbors required to classify, i.e., the parameter K is chosen as 5.

No		**2857**	**116**
Yes		**516**	**274**
Test	Predicted	No	Yes

The recall is **96.09%.**
The number of correctly classifying as 'yes' is **273.**

SUMMARY OF ALL MODELS

Model	**Accuracy**	**Recall**	**Correctly classified as 'yes'**
Logistic regression	79	99	71
Random forest	83	99	193
k-Nearest neighbors	83	96	273

CHOOSING THE MODEL

From the above table, the most balanced model, which provides the best recall, accuracy as well choose a good number of individuals to be contacted is the **Random Forest** model. So, this model is trained and extracted into a file called 'model.pkl' attached with the zip folder.

USING THE MODEL AND FURTHER ANALYSIS

All the files required attached in the zip folder. Below is a list of all files and how it can be used.

File name	**Usage**
Coursework _Summarization notebook.ipynb	To check how plots are visualized and it can be used to visualize any other new feature.
Coursework_Exploratory data analysis notebook.ipynb	To check how decision tree is drawn and play with different parameters if required.
Coursework_Model selection.ipynb	To check fitting data with any other model or use different parameters for existing model.
Coursework_Prediction.ipynb	Steps to use model and predict on datasets.
model.pkl	Model file that can be imported and used to classify new data points.

BUSINESS CASE RECOMMENDATIONS

N/LAB Enterprises is recommended to market their product “N/LAB Platinum Deposit” based on previous details of individuals being marketed to about a similar product.

The advantage is that, using the data and machine learning techniques, N/LAB can confidently get insights before marketing it to everyone which increases the cost. Since their strategy is to generate capital using this product, the chosen model helps in identifying highly potential individuals.

Although the risk involved is that no model is 100% accurate, and N/LAB might lose potential customers by choosing only those who are highly likely.

It is recommended to collect further data on newly named product, perform analysis on that and keep updating the model used for classifying, because rebranding a product can always have drastic impacts on how an individual views the product.