Post Graduate Diploma in Artificial Intelligence and Machine Learning

December 2024 Cohort

Leveraging Machine Learning to Improve Lead Quality for Bank-initiated Campaigns

ALFREDO C. SANCHEZ JR.

CAPSTONE PROJECT · DECEMBER 25, 2025

TABLE OF CONTENTS

I. Problem Understanding and Framing

When a bank initiates campaigns aimed at growing the business, it is usually met with an optimization problem on driving the highest-possible return on the budget allocated for the campaign. Put simply, given a budget, the campaign owner can only reach out to a finite subset from a pool of customers usually in the millions especially for retail products offered to individual clients.

Generally, the simplest way to determine who to offer would be via random selection out of all active and eligible (e.g., not in negative lists) customers, which is inefficient given the absence of thought and deliberate disregard for available information that can help with targeting. A better way is to tap the subject matter experts of the offer, usually the product owners who defined who their target market is and what makes the product relevant to that market. Data can then be used to create a rule-based selection of leads in order to ensure offer is relevant to the leads to achieve higher availment or conversion. In this capstone, a third option that leverages machine learning capabilities is explored with the goal of further improving the success of campaigns. At the end of the day, it is a return on investment business question: **how can we leverage data available to us about the customers' demographics, relationships, transactions, behavior, and interactions to deliver higher campaign ROI**.

In this capstone, an established dataset on bank marketing is used to illustrate how machine learning algorithms can help drive ROI. The problem will be approached two-fold: (1) create customer segmentation so it is possible for the banks to use different tactics per identified segment, and (2) develop a propensity model that will challenge the random and rule-based selection options. For the first one, K-means was used for the classification task; for the second, logistic regression, gradient boosting method, and XGBoost were considered then an ensuing hyperparameter tuning effort was done. For this second one, Accuracy and ROC AUC score are the main metrics to consider when making the choice. Precision, Recall, and F1-score were also looked at combined with qualitative considerations like interpretability and explainability which are both crucial particularly in a highly-regulated industry like banking.

Once the above were done, a comparison of the three (3) approaches—random, rules-based, and machine learning results—were done to determine the best approach using this sample dataset as illustration. Key performance assessment will be based on higher acceptance rate which will then be assessed against the cost. Naturally, there is an additional cost to employing machine learning capabilities so the **incremental lift arising from its employ should be greater than the corresponding incremental cost to develop and maintain the model**. A more detailed framework will be discussed in succeeding sections, particularly the one on campaign execution in the corporate communication portion.

II. Data Collection and Understanding

For the purpose of this capstone, Bank Marketing data—an established learning data set by Henrique Yamahata with source at https://archive.ics.uci.edu/ml/datasets/bank+marketing—will be used. With data and descriptions taken from the Kaggle website (link: Bank Marketing), the public is informed that these are data from a Portuguese banking institution. Initially, the chief goal of this dataset is to predict if a client will subscribe to a term deposit based on an availment/subscription agreement variable present in the data. In this case where there are information related to recent contact interactions on top of the available bank client data, the objective is to use what are relevant to predict the likelihood of acceptance/subscription. Below is the compiled data dictionary using a mix of what are readily available information in the site and some details added as assumptions in the data use specific to this capstone.

TABLE 1: *Bank Marketing Data Dictionary*

Variable	Type	List of Values	Description	Used in Capstone
Bank Client Data				
Age	Numeric	17 – 98	Assumed to be the age of the customer at time of data pull rounded down to the nearest number of years	Yes
Job	Categorical	'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown'	Classification of the type of job/employment status pf the customer. This assumes that a customer falls into one of the values in the list; noting that "unknown" is an option that is not to be confused with "others".	Yes
Marital	Categorical	'divorced', 'married', 'single', 'unknown'	Marital status of the customer; note on the data dictionary provided indicates that "widowed" is grouped with "divorced". Noting that "unknown" is an option.	Yes
Education	Categorical	'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown'	Educational attainment of the customer. Noting that "unknown" is an option.	Yes
Default	Categorical	'no', 'yes', 'unknown'	Indicator if the customer has a credit in default	Yes
Housing	Categorical	'no', 'yes', 'unknown'	Indicator if the customer has a housing loan	Yes
Loan	Categorical	'no', 'yes', 'unknown'	Indicator if the customer has a personal loan	Yes

Variable	Type	List of Values	Description	Used in Capstone
Customer Contact Interaction				
contact	Categorical	'cellular', 'telephone'	Channel used by the bank to communicate with the customer. Assumption is that it shows the most recent information in the bank's system or where the customer was successfully contacted.	Yes
month	Categorical	'jan', 'feb', 'mar', 'apr', 'may', 'jun', 'jul', 'aug', 'sep', 'oct', 'nov', 'dec'	Last contact month of the year represented by the first three letters of the name of the month.	No
day_of_week	Categorical	'mon', 'tue', 'wed', 'thu', 'fri'	Last contact day of the (work) week represented by the first three letters of the name of the day	No
duration	Numeric	0 - 4,918	last contact duration, in seconds. Capturing the important note from the provided data dictionary in verbatim: *This attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.*	No
Campaign-related Information				
campaign	Numeric	1 – 56	number of contacts performed during this campaign and for this client	No
pdays	Numeric	0 - 27; 999	number of days that passed after the client was last contacted from a previous campaign; note that 999 means client was not previously contacted	No
previous	Numeric	0 – 7	number of contacts performed before this campaign and for this client	Yes; transformed
poutcome	Categorical	'failure', 'nonexistent', 'success'	outcome of the previous marketing campaign	Yes
Other Information (e.g., Economic Data)				
emp.var.rate	Numeric	-3.4 - 1.4	employment variation rate - quarterly indicator	No
cons.price.idx	Numeric	92.201 - 94.767	consumer price index - monthly indicator	No
cons.conf.idx	Numeric	-50.8 - -26.9	consumer confidence index - monthly indicator	No
euribor3m	Numeric	0.634 - 5.045	euribor 3 month rate - daily indicator	No
nr.employed	Numeric	4,963.6 - 5,228.1	number of employees - quarterly indicator	No
Target/Predicted Variable				
Y	Binary	'yes','no'	indicates whether a client subscribed to a term deposit or not.	Yes

While a lot of information are made available in this dataset, certain fields were dropped given the objective of the capstone. The social and economic variables were removed to ensure the model is limited to data internally available especially as the author begins to think how such an approach can be applied to his current banking workplace which may not have the same information readily available. Also, the campaign data which may be related to seasonality were excluded for the reason that the goal is to have an always-on leads optimization initiative. The campaign info may be more useful in setting up the campaign calendar to know which time of the year take up spikes, but may be less useful if the goal is to establish propensity to accept/subscribe to the term deposit product at any given time.

Of the 41,188 entries in the dataset, there are no missing values; the 1,515 for pdays is expected given that the blank ones are those that were not previously contacted. See summary below:

TABLE 2: *Data Column Properties*

```
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 14 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        41188 non-null  int64
 1   job        41188 non-null  object
 2   marital    41188 non-null  object
 3   education  41188 non-null  object
 4   default    41188 non-null  object
 5   housing    41188 non-null  object
 6   loan       41188 non-null  object
 7   contact    41188 non-null  object
 8   duration   41188 non-null  int64
 9   campaign   41188 non-null  int64
 10  pdays      1515 non-null   float64
 11  previous   41188 non-null  int64
 12  poutcome   41188 non-null  object
 13  y          41188 non-null  int64
```

As a last note on the data, initial looks on univariate distributions and summary statistics show information reaching higher values which may have an outlier impact to the model, which is why some data-handling steps were carried out to ensure such risk is avoided. This and other data transformation and handling steps are discussed in the next section.

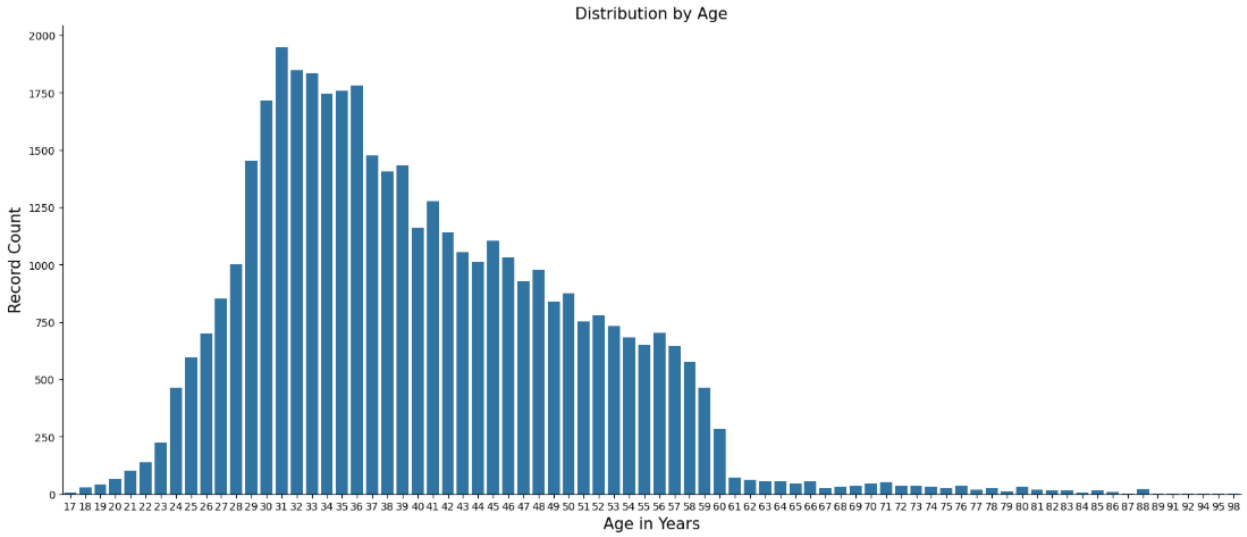
III. Data Preprocessing, Applied EDA and Feature Engineering

As mentioned in the previous section, one of the reasons why this sample dataset was chosen is because of its high-usability and lack of missing values. Hence, minimal data wrangling is needed for the purposes of this capstone.

First thing the author looked at is the balance of the target variable, y. Looking at the distribution, only 11% accepted the term deposit leading to an unbalanced set. No need to change this, but the author made sure to employ balancing logic in the model build.

For age, it appears to be positively skewed as can be seen in Graph 1 below. Nevertheless, given the concentration of datapoints within the 23-60 range, practicality considerations lead the author to believe it should be fine to retain as-is.

GRAPH 1: *Number of Customers by Age in Years*

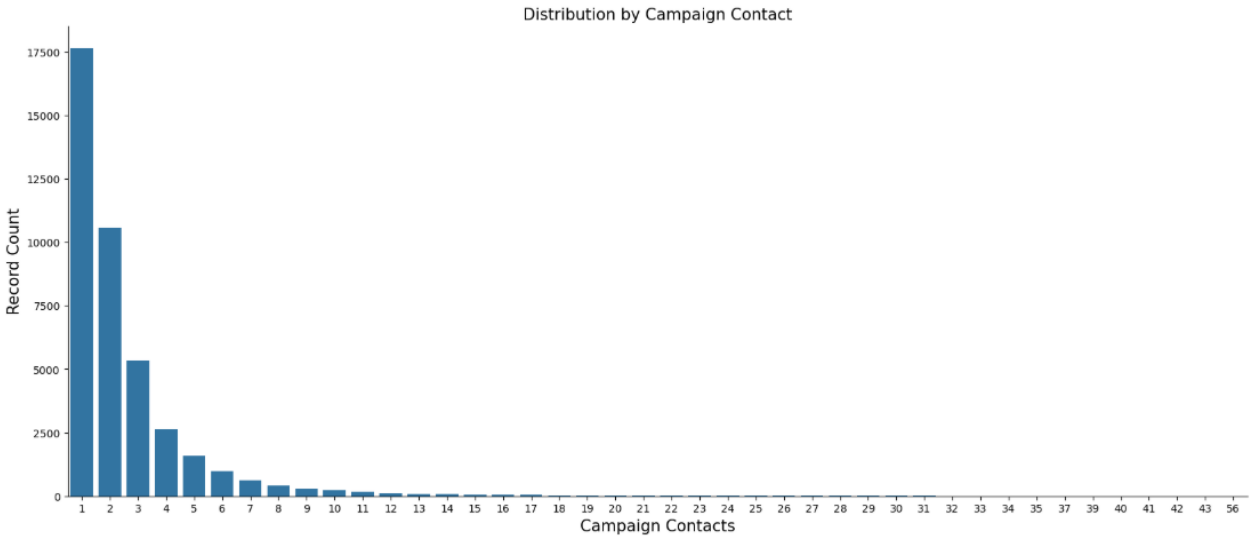


On the other hand, looking at the campaign contact information in Graph 2 in the following page, it is apparent that there are a lot of low-count records for those beyond 5. Given this, the author believe it is best to impose a ceiling of 5 given that the objective of this column is simply to establish the number of contacts done for the campaign. Graph 3 in the succeeding page shows the effect of this ceiling imposition

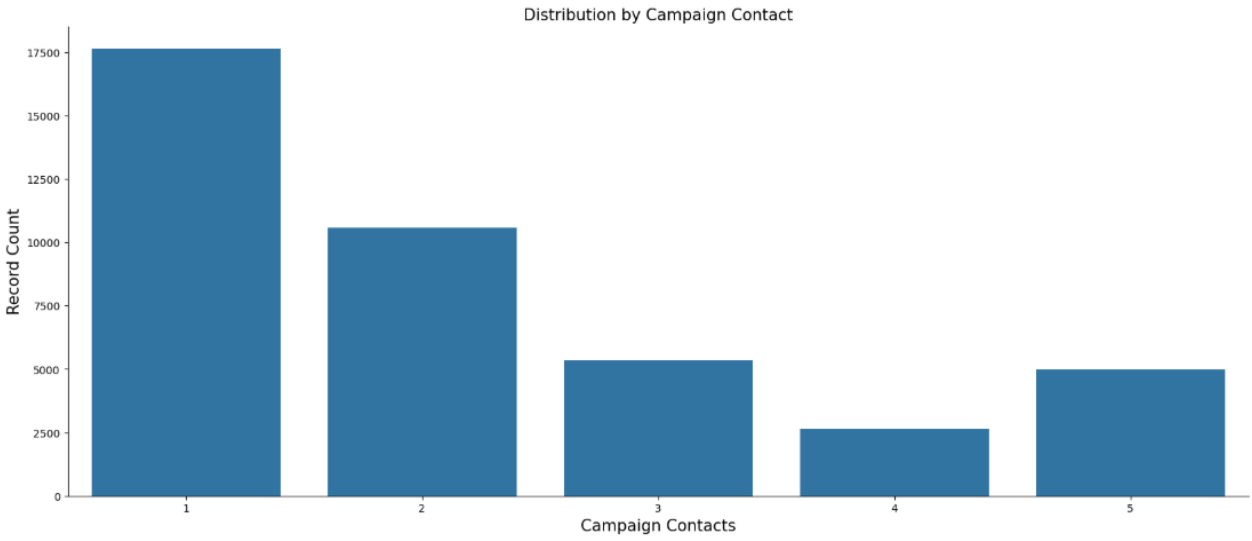
Nevertheless, while this exercise of ceiling imposition is done, it is worth noting that this variable was eventually excluded from the model given it's dynamic of only being known once the customer is *already* part of the leads list. Note that the objective of this exercise is to optimize lead quality, which means that all information used for this activity should be data that are available *before* the leads list is generated.

Similar to the note on campaign timing being more useful in creating campaign calendars, this information on the number campaign contacts may be more useful in determining the appropriate call intensity when contacting customers that are eligible for the offer.

GRAPH 2: *Number of Customers by Number of Contacts Made During the Campaign*

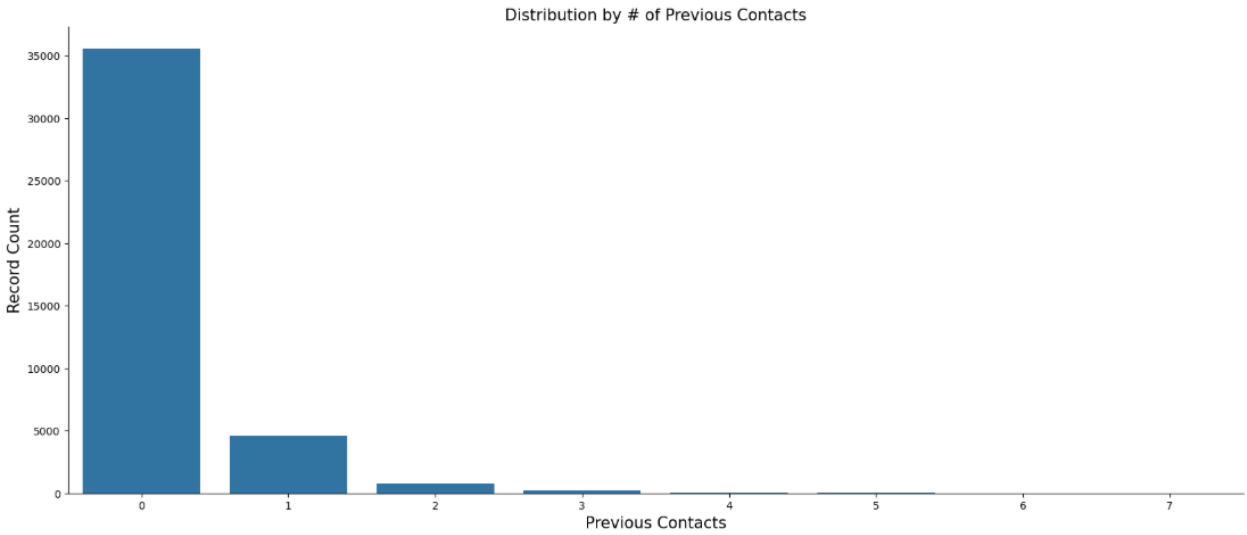


GRAPH 3: *Number of Customers by Number of Contacts Made (Max = 5) During the Campaign*



In terms of the information on previous contacts, Graph 4 in the following page shows the initial distribution of the number of previous contacts made to the customer. Note that the number of customers that have at least one contact is quite low, which is why it was decided to convert this variable to a binary variable where those that register zero are labeled as without prior contact. After this transformation, it is noted that 86% of the customer base were not contacted previously.

GRAPH 4: *Number of Customers by Number of Previous Contacts Made*



To do further Exploratory Data Analysis, frequency distributions were looked into to see how the accept/subscription rate differ by the different categorical variables as data cuts. See the following tables for reference. Given that the yeses for acceptance is at 11%, groups with acceptance rate >11% are highlighted.

TABLE 3: *Acceptance Rate by Job Status*

y	0	1
job		
admin.	**87.03**	**12.97**
blue-collar	93.11	6.89
entrepreneur	91.48	8.52
housemaid	90.00	10.00
management	**88.78**	**11.22**
retired	**74.77**	**25.23**
self-employed	89.51	10.49
services	91.86	8.14
student	**68.57**	**31.43**
technician	89.17	10.83
unemployed	**85.80**	**14.20**
unknown	**88.79**	**11.21**

Note that while student, unemployed, and unknown register higher-than-average accept rates, they have low bases in the sample so need to interpret with caution.

TABLE 4: *Acceptance Rate by Marital Status*

y	0	1
marital		
divorced	89.68	10.32
married	89.84	10.16
single	**86.00**	**14.00**
unknown	**85.00**	**15.00**

Note that unknown registers a higher-than-average accept rate, it has a low base in the sample so need to interpret with caution.

TABLE 5: *Acceptance Rate by Educational Attainment*

y	0	1
education		
basic.4y	89.75	10.25
basic.6y	91.80	8.20
basic.9y	92.18	7.82
high.school	89.16	10.84
illiterate	**77.78**	**22.22**
professional.course	**88.65**	**11.35**
university.degree	**86.28**	**13.72**
unknown	**85.50**	**14.50**

Note that while illiterate registers a higher-than-average accept rate, it has a low base in the sample so need to interpret with caution.

TABLE 6: *Acceptance Rate by Whether or Not Client Has Defaulted*

y	0	1
default		
no	**87.12**	**12.88**
unknown	94.85	5.15
yes	100.00	0.00

TABLE 7: *Acceptance Rate by Housing Loan Availment Status*

y	0	1
housing		
no	89.12	10.88
unknown	89.19	10.81
yes	**88.38**	**11.62**

TABLE 8: *Acceptance Rate by Personal Loan Availment Status*

y	0	1
loan		
no	**88.66**	**11.34**
unknown	89.19	10.81
yes	89.07	10.93

TABLE 9: *Acceptance Rate by Contact Channel*

y	0	1
contact		
cellular	**85.26**	**14.74**
telephone	94.77	5.23

TABLE 10: *Acceptance Rate by Previous Campaign Outcome*

y	0	1
poutcome		
failure	**85.77**	**14.23**
nonexistent	91.17	8.83
success	**34.89**	**65.11**

Note that while failure and success register higher-than-average accept rates, they have low bases in the sample so need to interpret with caution. This is an effect of these customers likely not being part of prior campaigns.

TABLE 11: *Acceptance Rate by Whether or Not Customer was Previously Contacted*

y	0	1
prior_contact		
no	91.17	8.83
yes	**73.35**	**26.65**

From the above information which is merely a check on variable-level information on its impact to accept rate, this would be a good basis for establishing a rules-based logic in selecting customers. For the purpose of this capstone, the rules-based challenger will be the following:

- Age must be between 30 and 60 and the contact channel is cellular, or
- Prior campaign outcome was a success

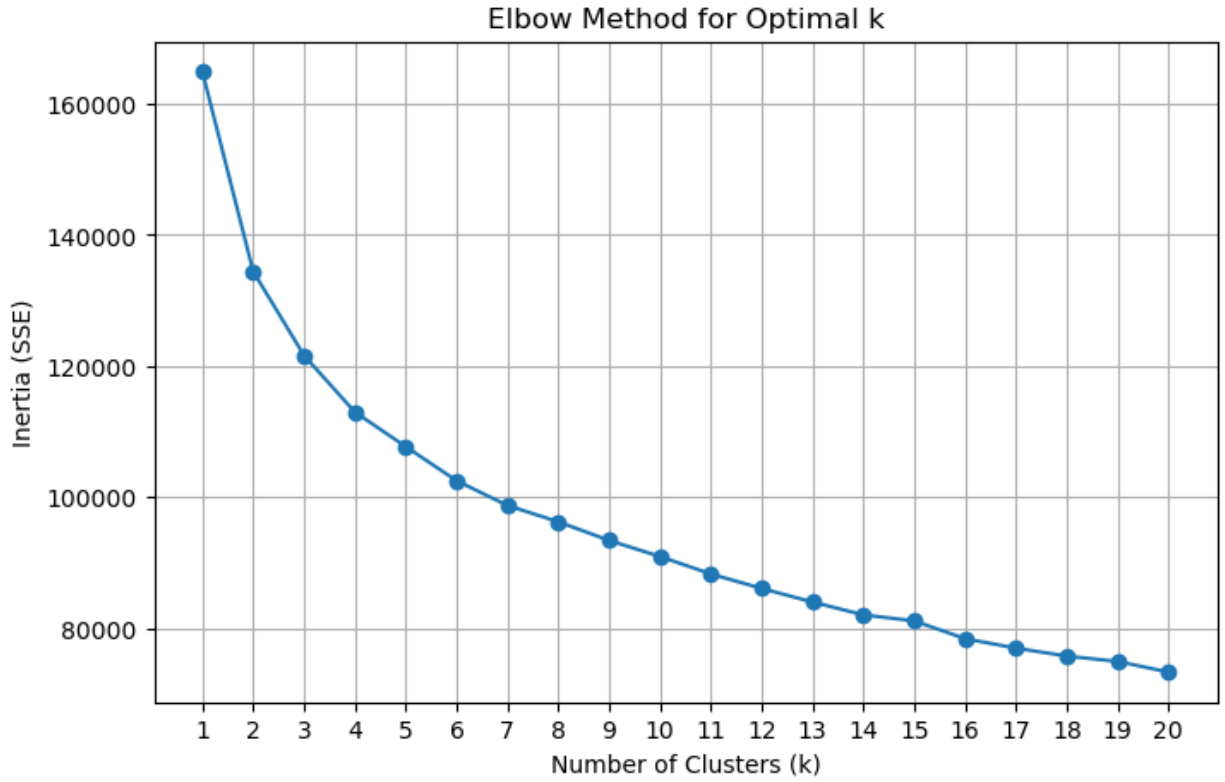
The succeeding section will discuss how K-means clustering was used to create customer segmentation and then how a final model to determine propensity to subscribe to the term deposit was achieved.

IV. Model Implementation

In terms of creating customer segmentation, the author wanted to limit it to customer characteristics that are readily observable: age, job status, marital status, educational attainment, loan product availments. Standard scaler was applied to age while one hot encoding was used for the categorical variables.

To know the number of clusters to use, SSE was plotted against the evaluated possible number of clusters, *k*. Initially, only up to 10 clusters were assessed, but this was expanded to 20 given the absence of an obvious elbow. Looking at Graph 5 below—even if there still is no clear elbow—there seems to be a slight stabilization happening around $k = 6$, which is why this is the number of clusters chosen. Besides, too many clusters can lead to confusion and over-segmentation.

GRAPH 5: *Inertia (SSE) by the Number of Clusters (k)*



With the choice of six (6) segments, the summary interpretation is shown in Table 12 in the succeeding page. From this table is clear that two groups are comprised of young and single customers (segments 0 and 4), two groups are comprised by married customers in their mid-30s (segments 3 and 5), and two groups are in their late 40s (segments 1 and 2). These 3 subgroups are split into whether or not the customers availed of a housing loan, which leads to 6 segments.

TABLE 12: *Customer Segment Interpretation*

segment	age	job	marital	education	housing	loan
0	30	admin.	single	university.degree	no	no
1	48	blue-collar	married	university.degree	yes	no
2	46	blue-collar	married	university.degree	no	no
3	36	blue-collar	married	university.degree	yes	no
4	31	admin.	single	university.degree	yes	no
5	36	blue-collar	married	university.degree	no	no

The segment variable from this exercise is used in the following propensity modeling.

In this modeling exercise, three options were evaluated: Logistic Regression, Gradient Boosting Method, and XGBoost. In Table 13 below, it is clear that the best method using the metrics considered is Gradient Boosting.

Table 13: *Model Performance Comparison*

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Gradient Boosting	**0.896010**	**0.630170**	**0.186063**	**0.287299**	**0.734525**
Logistic Regression	0.896981	0.651399	0.183908	0.286835	0.722916
XGBoost	0.894068	0.596288	0.184626	0.281953	0.714449

However, given that the differences between the models are low, the author chose to stick to using Logistic Regression for better explainability and interpretability, which are critical for highly-regulated industries like banking. In the Philippines, specific to the mandate of the Bangko Sentral ng Pilipinas, it is imperative that it is clear to the customers how their information is used especially in determining their eligibility for offers. While Gradient Boosting might be better for models requiring more complex and flexible approaches to learning, it loses to Logistic Regression on the two aforementioned critical qualities.

Given the choice to stick to Logistic Regression, the author checked if it is worth updating the parameters to improve model performance. Graph 6 on the following page shows the hyperparameter tuning results while Table 14 shows the accuracy and performance reports of the logistic regression models using the different parameters

GRAPH 6: *Hyperparameter Tuning Results*

```
Best parameters found (Grid Search): {'C': 10, 'penalty': 'l1', 'solver': 'liblinear'}
Best CV score (Grid Search): 0.6784707217094692
Best parameters found (Random Search): {'C': 1.7477481403880757, 'penalty': 'l1', 'solver': 'liblinear'}
Best CV score (Random Search): 0.6785054078003466
```

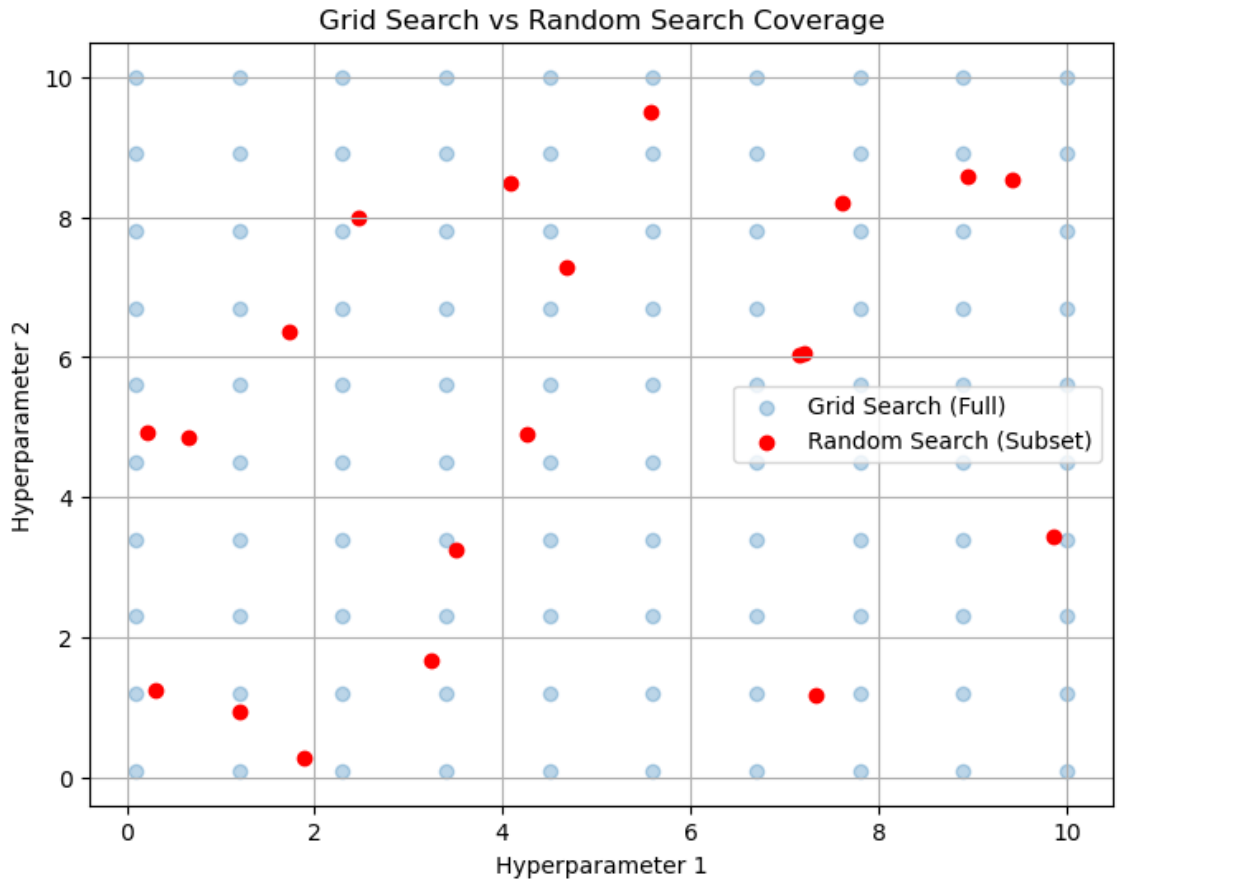


TABLE 14: *Logistic Regression Models Comparison*

```
Accuracy Score (Original): 0.683094602249737
Classification Report (Original):
              precision    recall  f1-score   support

           0       0.94      0.69      0.79     10965
           1       0.21      0.63      0.31      1392

    accuracy                           0.68     12357
   macro avg       0.57      0.66      0.55     12357
weighted avg       0.85      0.68      0.74     12357
```

```
Accuracy Score (Grid Search): 0.6823662701302905
Classification Report (Grid Search):
              precision    recall  f1-score   support

           0       0.94      0.69      0.79     10965
           1       0.20      0.63      0.31      1392

    accuracy                           0.68     12357
   macro avg       0.57      0.66      0.55     12357
weighted avg       0.85      0.68      0.74     12357

Accuracy Score (Random Search): 0.6824471959213402
Classification Report (Random Search):
              precision    recall  f1-score   support

           0       0.94      0.69      0.79     10965
           1       0.20      0.63      0.31      1392

    accuracy                           0.68     12357
   macro avg       0.57      0.66      0.55     12357
weighted avg       0.85      0.68      0.74     12357
```

Given the highest accuracy shown by the original model, the author decided to stick with the first, unadjusted result. With its ROC AUC score of 0.7229, it is generally considered acceptable especially for banking use cases on propensity modeling. Moreover, since this is about assessing campaign customer leads and not something more accuracy-demanding like fraud use cases, this model should be acceptable for the objectives of this capstone.

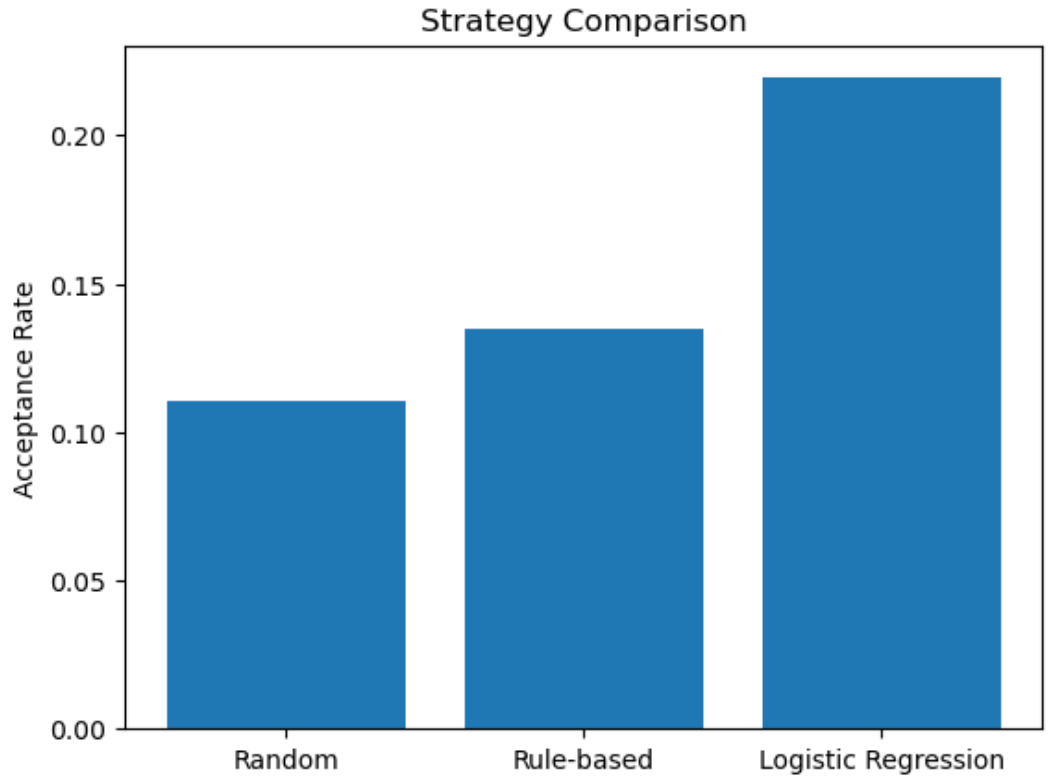
The details of the model are laid out in the technical presentation with the main variables used being: age, job status, marital status, educational attainment, whether or not the client has defaulted, prior outcome success, and loan product availments.

In the simulation to compare the 3 approaches: random, rules-based logic, and logistic regression, a contact rate of 30% is assumed.

As seen in Graph 7 in the following page, random's accept rate of 11% was improved by a rule-based approach but only by 2.5 percentage points to 13.5%. The significant jump is via Logistic Regression which sees an accept rate of 22% which is twice the success of the random approach. It is clear that there is significant benefit to employing machine learning to optimize lead quality. Nonetheless, this should definitely be viewed vis-à-vis the cost of developing and maintaining the models. It is a must that the lift enjoyed in revenue from the accept rate doubling is higher than the cost to deploy and employ such a model.

GRAPH 7: *Approach Comparison (Random vs. Rule-based vs. Logistic Regression)*

	strategy	contacts	accepts	accept_rate
0	Random	3707	409	0.110332
1	Rule-based	3707	500	0.134880
2	Logistic Regression	3707	813	0.219315



V. Critical Thinking → Ethical AI and Bias Auditing

In the case of this example, there is not much bias that needed to be addressed aside from the need to balance. The ethical consideration as well as bias considerations will likely stem from ensuring that the eligibility is free from any discriminatory or unfair qualifications (or perhaps singling out a specific subgroup). In this particular example, there is not much of a concern as an eyeball of parameters show that there is minimal to no source of adverse or favorable bias. It may be unfortunate in the case of showcasing specific to this capstone, but suffice it to say that part of the beauty of choosing this dataset is its cleanliness and lack of tendency to lead to biased outcomes.

Nevertheless, it is critical to ensure that bias checks continue to be monitored so that corrective action can be done asap especially the given mentioned strictness of regulations governing the banking industry. One good way to ensure is to parameterize penalties so that there is a clear safeguard against bias that may introduce itself later on.

It is important to watch for this to ensure fairness in performance across different demographic groups, whether favorably or adversely.

VI. Final Presentation and Communication

Two presentations are prepared and attached as supporting documents. For the purposes of discussion in this main capstone document, the author will share details on the contents of the respective presentations.

For the technical presentation, kindly refer to the attached ***Alfredo Sanchez Jr_Capstone_2_Technical Presentation.html***. This version is basically the Jupyter notebook with the outputs; something that can be used to discuss with people proficient on the technical aspects and can understand and appreciate coded details.

For the strategic presentation for senior leadership, kindly refer to the attached Powerpoint presentation, ***Alfredo Sanchez Jr_Capstone_3_Strategic Presentation.ppt***. In this business deck, the ROI conversation happens as well as the considerations that the business needs to care for in executing against the leads list, which may not be immediately apparent in the model build. For example, concurrent eligibility in other offers as some may take priority. Another would be the employment of a champion-challenger strategy to not only prove that using Logistic Regression is beneficial but to have a clear capture of the lift vis-à-vis other options like a rule-based approach. A framework to assess is shown in one of the slides in the business presentation.

In the business application and execution, the campaign and product owners will be able to address the main problem of which customers to reach out to given the limited resources of budget, time, and manpower to maximize the expected yield via acceptance or subscription rate in this specific capstone example. This result can be extended to other campaigns and products under the banking industry.

VII. GitHub Profile and Upload

Kindly visit the link, alfsanchez/AIML-Capstone-202512: This repository is created to house all data and documents used to complete the Post Graduate Diploma on Artificial Intelligence and Machine Learning

This public repository houses all the required data and documents used for this capstone.

VIII. Deployment and MLOps (Optional)

While there was an attempt to do this section, the author only reached until the setting up of the codes for Flask deployment as can be seen in the Technical Presentation. No need to check on this as this optional portion was not completed.