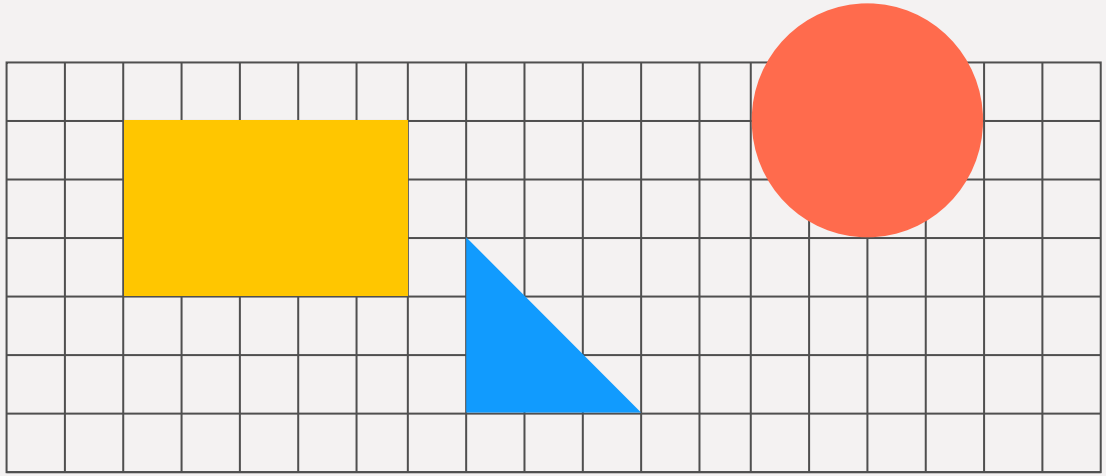
Predicting Bank Marketing Success



By Ravin Kabra, Greg Adler, Kushal Patel, Abdifatah Abdi, Carter Jensen

Our Data

The Bank Marketing dataset, obtained from the UCI Machine Learning Repository, contains data related to direct marketing campaigns (phone calls) of a Portuguese banking institution.

The dataset consists of various features, including:

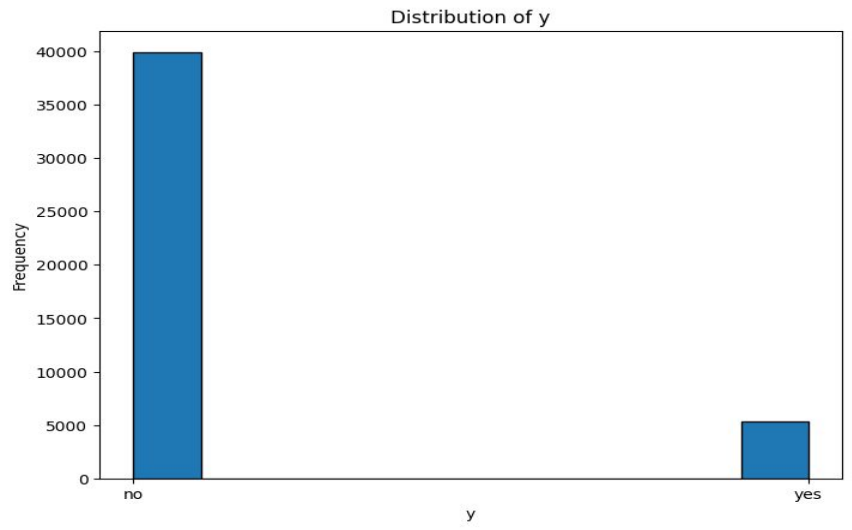
- *Call duration*
- *Education level*
- *Bank balance*
- *Past outcomes*

And target (y), whether or not the client subscribed to a term deposit

Introduction

Our Goal

Discover patterns in the subscribers and nonsubscribers to more effectively and efficiently allocate marketing efforts and resources



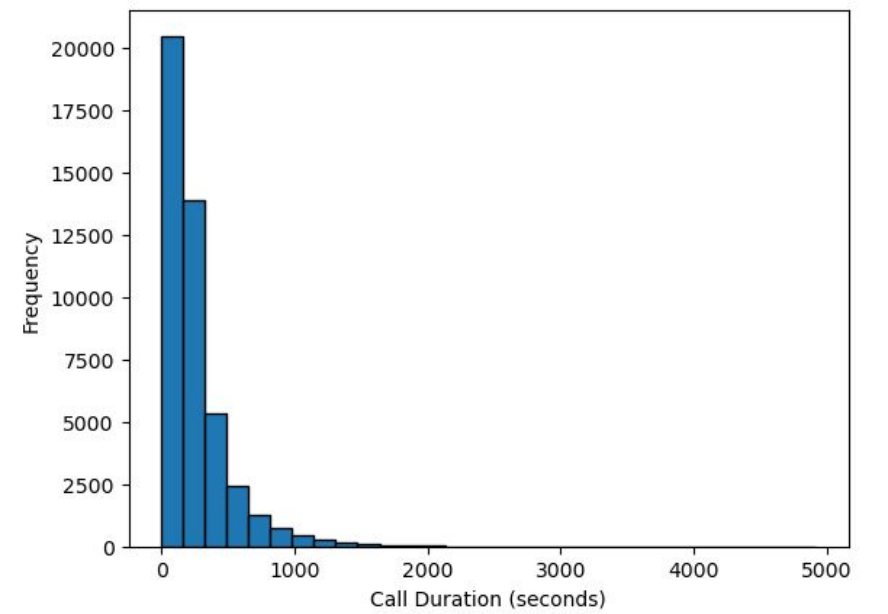
- The proportion of y = 1 (proportion of people who subscribed for a term deposit) in our data is 0.12
- With such a low proportion of subscribers, it will be important that our model captures as many as possible, so high recall was a main focus of ours.

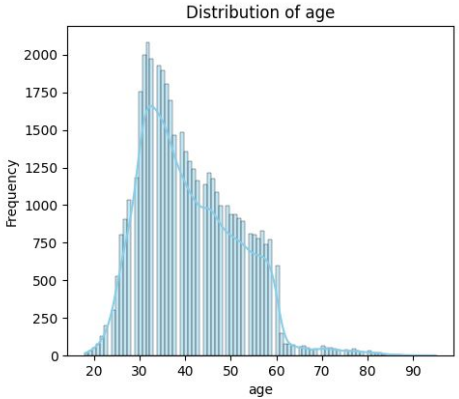
Pre-Processing, Methods and Approaches

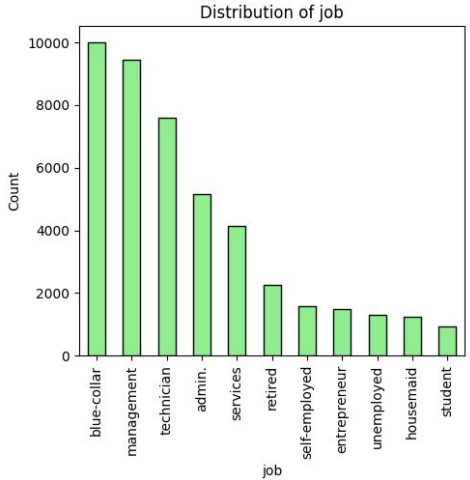
Based on the distribution of our data, it would be apt to use standardization vs normalization due to the presence of outliers since normalization is more sensitive to outliers.

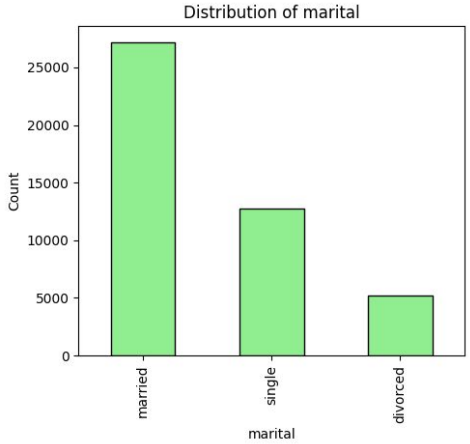
Based on our data and goal, it would be apt to approach this problem with a classification model. We wish to see how well can the model support data driven decision making for our marketing endeavours to ensure that the return on marketing investment is optimal

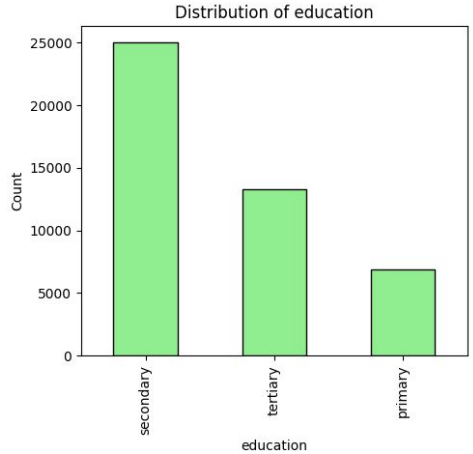
We shall proceed with logistic regression based model as well as using K means clustering to verify and validate the model results to get better marketing insights

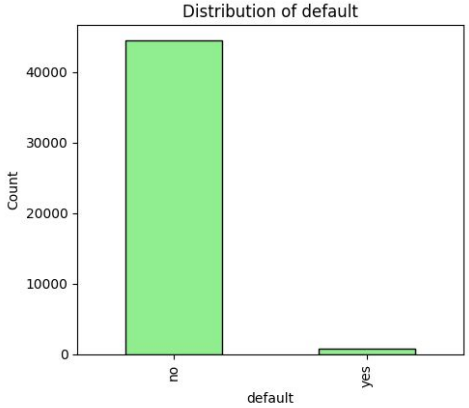


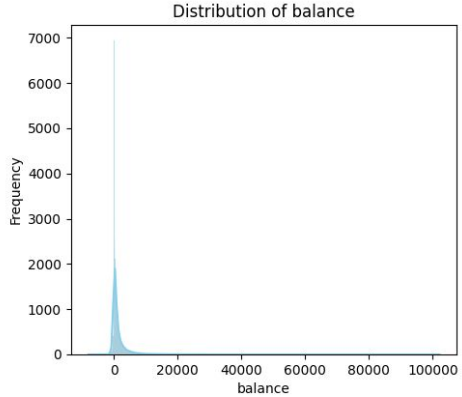

Distribution of age
Frequency
age

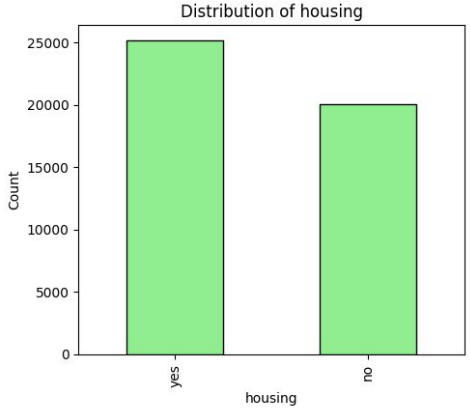

Distribution of job
Count
blue-collar
management
technician
admin.
services
retired
self-employed
entrepreneur
unemployed
housemaid
student
job

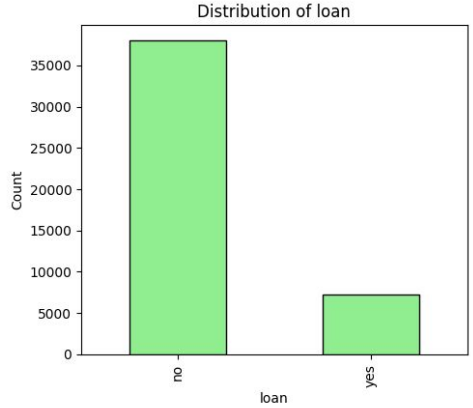

Distribution of marital
Count
married
single
divorced
marital

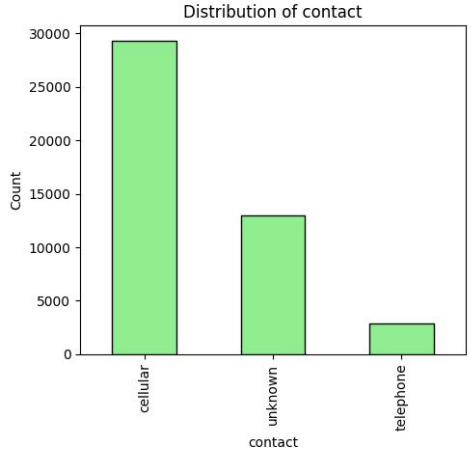

Distribution of education
Count
secondary
tertiary
primary
education

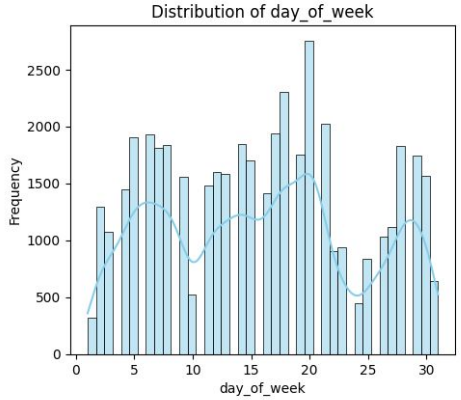

Distribution of default
Count
no
yes
default

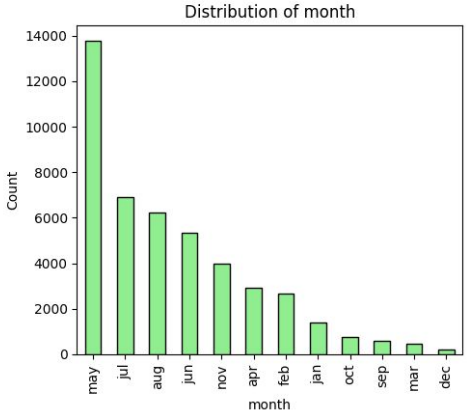

Distribution of balance
Frequency
balance

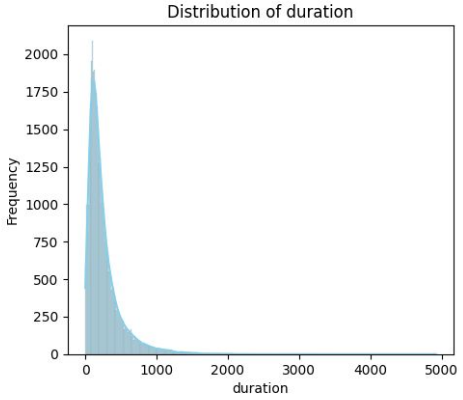

Distribution of housing
Count
yes
no
housing

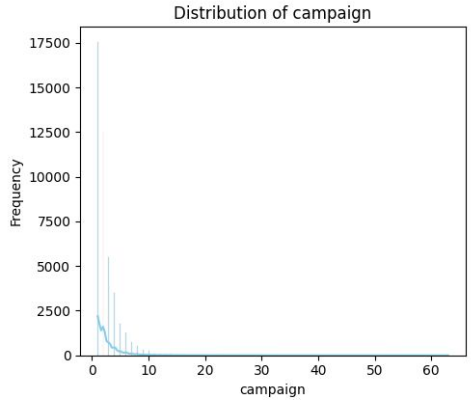

Distribution of loan
Count
no
yes
loan

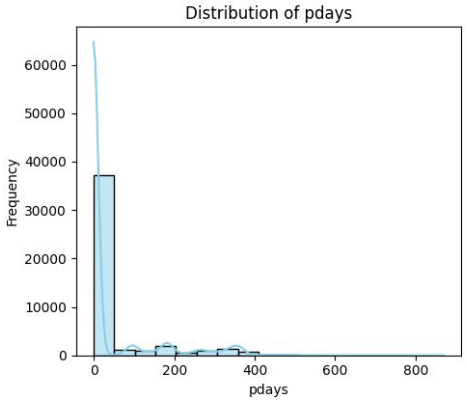

Distribution of contact
Count
contact
cellular
unknown
telephone

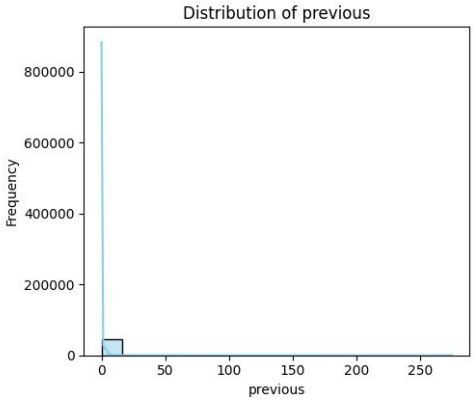

Distribution of day_of_week
Frequency
day_of_week

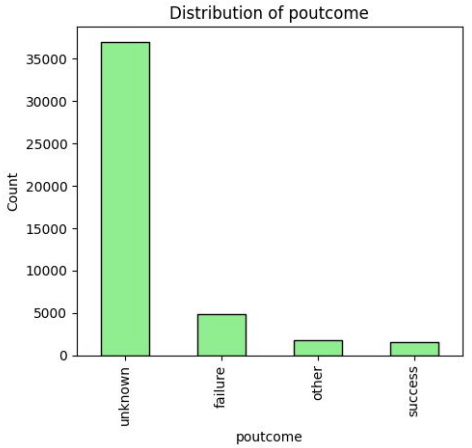

Distribution of month
Count
month
may
jul
aug
jun
nov
apr
feb
jan
oct
sep
mar
dec


Distribution of duration
Frequency
duration


Distribution of campaign
Frequency
campaign


Distribution of pdays
Frequency
pdays


Distribution of previous
Frequency
previous


Distribution of poutcome
Count
poutcome
unknown
failure
other
success

Top 20 features by coef

Feature	Coefficient
poutcome_success	2.468253
month_mar	1.6028
contact_nan	-1.41942
education_primary	-1.380474
housing_yes	-1.240856
month_jan	-1.183433
education_nan	-1.179516
education_secondary	-1.173794
duration	1.086122
job_housemaid	-1.033131
education_tertiary	-1.004767
month_sep	0.961138
month_oct	0.935699
month_dec	0.863161
job_self-employed	-0.843446
job_entrepreneur	-0.82463
month_nov	-0.793275
job_blue-collar	-0.788871
month_jul	-0.74521
job_services	-0.706993

Summary of Feature Interpretations

The following key features significantly impact the likelihood of a client subscribing:

Positive Impact (Increases Subscription Likelihood)

- `poutcome_success`: Clients with successful outcomes in previous campaigns are much more likely to subscribe.
- `month_mar`, `month_sep`, `month_oct`, `month_dec`: Clients contacted in these months are more likely to subscribe compared to other months.
- `duration`: Longer call durations strongly correlate with higher subscription rates, emphasizing meaningful client interactions.

Negative Impact (Decreases Subscription Likelihood)

- `education_primary` **and** `education_secondary`: Clients with primary or secondary education are less likely to subscribe than those with tertiary education.
- `housing_yes`: Clients with housing loans are less likely to subscribe, possibly due to financial constraints.
- `job_blue-collar`, `job_housemaid`, `job_self-employed`, `job_entrepreneur`, `job_services`: Clients in these professions are less likely to subscribe, reflecting financial priorities or constraints.
- `contact_nan`: Missing contact information is associated with lower subscription likelihood.
- `month_jan`, `month_nov`, `month_jul`: Clients contacted in these months are less likely to subscribe compared to others.

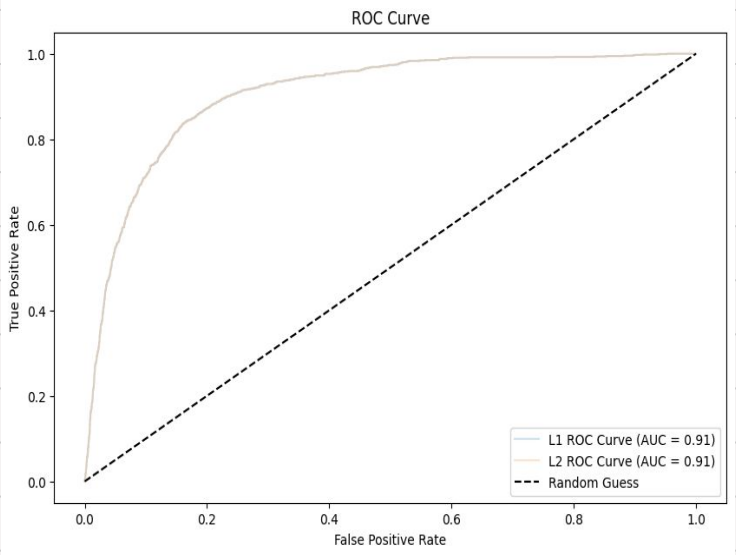
Key Insights

- **Successful Past Interactions Matter**: Prior campaign success is the strongest positive indicator.
- **Call Timing is Crucial**: Subscription likelihood varies significantly across months, with March, September, and October showing higher success rates.
- **Call Duration is Key**: Longer calls are highly correlated with subscription success.
- **Focus on Educated Clients**: Clients with tertiary education show better engagement and conversion rates.
- **Occupational Differences**: Certain job categories (e.g., blue-collar workers, housemaids) are less likely to subscribe, requiring tailored approaches.

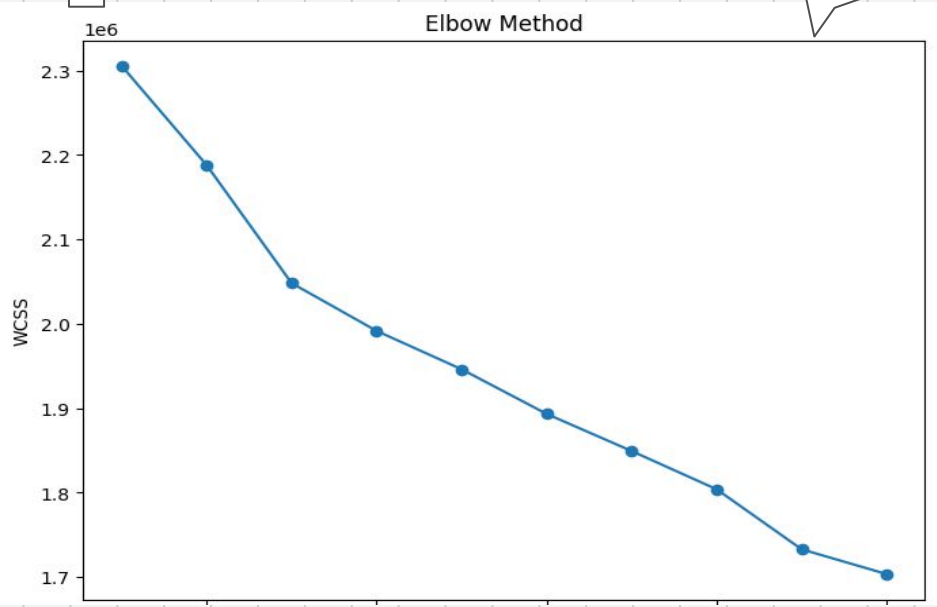
Logistic Regression

Based on grid search CV with Micro F1 scoring since our data is imbalanced, the scoring for L1 and L2 respectively is as follows. Since L1 and L2 have highly similar scores, it would be preferred to use L1 regularization as LASSO would have higher penalties and emphasize more on certain features over other. A stronger coefficient assignment shall help us focus are marketing efforts more on to those as opposed to a range of less strong features. Since we need to specialize marketing to classes and demographics and not get overall data trends, we should use L1.

Metric	L1 (Class 0)	L1 (Class 1)	Maco Avg (L1)	Weighted Average (L1)	L2 (Class 0)	L2 (Class 1)	Macro Avg (L2)	Weighted Avg. (L2)
Precision	0.92	0.65	0.78	0.89	0.92	0.65	0.78	0.89
Recall	0.97	0.35	0.66	0.90	0.97	0.35	0.66	0.90
F1-Score	0.95	0.45	0.70	0.89	0.95	0.45	0.70	0.89
Support	7985	1058	9043	9043	7985	1058	9043	9043
Accuracy			0.90				0.90	



Clustering



By using three clusters, we were able to optimize the trade-off between model performance and simplicity.

- **Cluster 0: First-Time Targets**
    a. **Key Characteristics**:
        i. Slightly positive average **age** and **balance**: Indicates a mix of older and moderately wealthy customers.
        ii. `day_of_week`: Positive value indicates higher engagement on specific days of the week.
        iii. `pdays` **and** `previous`: Negative values indicate this group was less engaged in prior campaigns and has longer gaps since last contact.
        iv. `poutcome_nan = 1.0`: All clients in this cluster have no outcome from previous campaigns, meaning they were not previously targeted or their outcome was not recorded.
    b. **Marketing Implications**:
        i. **Opportunities for First-Time Engagement**: This cluster may represent untapped potential, as these clients lack previous campaign outcomes.
        ii. **Focus on Awareness**: Since prior contact is minimal, awareness campaigns introducing products/services might be most effective.
- **Cluster 1: Previous, higher-earning Targets (Custer 1):**
    - **Key Characteristics**:
        - Higher **balance**: Indicates clients with better financial standing compared to other clusters.
        - Positive `previous` **(1.127)**: Clients in this cluster have been engaged frequently in past campaigns.
        - `poutcome_success = 0.183`: Shows a moderate success rate in previous campaigns.
        - `month_may`: High value indicates significant engagement during May campaigns.
        - `poutcome_failure = 0.594`: Many clients in this cluster had failed outcomes in previous campaigns.
    - **Marketing Implications**:
        - **Re-Engagement Needed**: Since many previous attempts have failed, tailor messaging or offer new incentives to convert this segment.
        - **Timing is Key**: May campaigns seem particularly relevant for this group.
        - **Potential for High ROI**: These clients have shown some engagement, making them easier to convert with optimized strategies.
- **Cluster 2: Previous, lower-earning, Targets**
    - **Key Characteristics**:
        - Slightly younger clients (**age = -0.077**) with lower **balance**: Indicates a financially constrained demographic.
        - Negative `duration`: Shorter call durations indicate limited engagement during calls.
        - `job_blue-collar` **= 0.325**: Blue-collar workers dominate this cluster.
        - `month_may`: Very high value indicates most of this group was contacted in May.
        - `poutcome_nan = 1.0`: Like Cluster 0, there are no recorded outcomes from previous campaigns.
    - **Marketing Implications**:
        - **Focus on Low-Cost Products**: Financial constraints suggest promotions or budget-friendly offers might resonate more with this group.
        - **Tailored Outreach**: Engage in shorter, focused calls to improve efficiency.
        - **New Target Audience**: May campaigns might be effective, but this group appears underutilized, requiring new approaches.

Intersection of Clustering Insights and Logistic Regression Coefficients

The clustering analysis and logistic regression coefficients provide complementary perspectives on the Bank Marketing dataset. While clustering groups clients into distinct segments for targeted strategies, logistic regression identifies features most strongly associated with term deposit subscription. Here's how the insights from both methods intersect:

Key Overlaps

Duration of Calls:

- **Logistic Regression Insight**: A strong positive coefficient for `duration` indicates that longer calls significantly increase the likelihood of subscription.
- **Clustering Insight**: Cluster 2 shows negative `duration`, indicating shorter calls, and this aligns with their lower likelihood of subscription.
- **Intersection**:
  - Focus on longer, meaningful interactions for clusters like Cluster 2, where shorter calls are common.
  - For Cluster 1, which has moderately positive `duration`, ensure call scripts are engaging and persuasive.

Month of Contact:

- **Logistic Regression Insight**: Certain months (e.g., `month_mar`, `month_sep`, `month_oct`) have positive coefficients, while others (e.g., `month_jan`, `month_nov`) are negative.
- **Clustering Insight**:
  - Cluster 1 is highly active in **May**, with some success but also high failure rates, suggesting further optimization is needed.
  - Cluster 2 is predominantly contacted in **May** but lacks prior campaign outcomes, representing untapped potential.
- **Intersection**:
  - Timing matters: Optimize campaign timing by targeting specific months where engagement and success rates are higher.
  - Leverage positive months (e.g., March, September, October) for high-priority clients in Clusters 0 and 1.

1. **Job Type**:
   - **Logistic Regression Insight**: `job_blue-collar`, `job_housemaid`, `job_self-employed`, and `job_services` have negative coefficients, indicating lower likelihoods of subscription.
   - **Clustering Insight**:
     - Cluster 2 is dominated by **blue-collar workers**, with shorter calls and lower financial balances.
   - **Intersection**:
     - Avoid investing heavily in lower-converting job segments like blue-collar workers or housemaids (Cluster 2).
     - For these groups, emphasize affordable options or alternative products.
2. **Poutcome (Previous Campaign Outcomes)**:
   - **Logistic Regression Insight**: `poutcome_success` has the highest positive coefficient, highlighting the importance of successful past interactions.
   - **Clustering Insight**:
     - Cluster 1 includes a moderate percentage of successful outcomes (`poutcome_success = 0.183`) and a significant number of failures (`poutcome_failure = 0.594`), indicating potential for improved re-engagement.
     - Clusters 0 and 2 have `poutcome_nan = 1.0`, meaning these clients lack prior campaign outcomes.
   - **Intersection**:
     - For Cluster 1, re-engage clients with a history of failed outcomes using revised offers or improved targeting.
     - For Clusters 0 and 2, implement awareness campaigns to engage clients with no previous outcomes.
3. **Education**:
   - **Logistic Regression Insight**: `education_primary` and `education_secondary` have negative coefficients, while `education_tertiary` has a weaker negative impact.
   - **Clustering Insight**:
     - Education plays a role in Cluster 0 (mixed education levels) and Cluster 2 (likely dominated by lower education levels).
   - **Intersection**:
     - Focus on clusters with higher education levels (e.g., tertiary) for premium product promotions.
     - For lower-education clusters, ensure messaging is simple, clear, and incentive-driven.

Key Differences

1. **Feature Weighting**:
    - Logistic regression provides a global view of feature importance across the dataset, while clustering focuses on segment-specific characteristics.
    - For example, `housing_yes` is globally significant (negative coefficient in regression) but may vary in importance across clusters.
2. **Granularity**:
    - Clustering creates actionable groups (e.g., Cluster 1 for re-engagement, Cluster 2 for awareness campaigns).
    - Logistic regression does not group clients but identifies universal trends.

Actionable Insights

- **Combine Logistic Regression and Clustering**:
    - Use logistic regression coefficients to identify high-impact features.
    - Apply clustering to segment clients and create tailored strategies for each group based on these features.
- **Examples**:
    - **Cluster 0 (Untapped Potential)**:
        - Prioritize outreach to clients with no prior outcomes (`poutcome_nan`) in positive months (e.g., March, September).
        - Leverage long call durations for meaningful engagement.
    - **Cluster 1 (Re-Engagement Candidates)**:
        - Focus on re-engaging clients with past failed outcomes using personalized offers.
        - Optimize campaigns for May, ensuring improved messaging or incentives.
    - **Cluster 2 (Financially Constrained)**:
        - Offer budget-friendly products to blue-collar workers.
        - Avoid high-investment strategies in January or November, as these months are less effective.

NEXT STEPS

1. Combine Clustering and Logistic Regression Insights

- **Segment-Based Targeting**: Use clustering results to tailor marketing strategies for each segment (e.g., re-engagement for Cluster 1, awareness campaigns for Cluster 0).
- **Feature-Specific Focus**: Apply logistic regression insights to prioritize impactful features like call duration, month of contact, and prior campaign outcomes.

2. Optimize Campaign Timing

- **Leverage High-Impact Months**: Focus efforts on months with positive coefficients (e.g., March, September, October) for all clusters.
- **Reduce Efforts in Low-Impact Months**: Minimize marketing spend during months with negative coefficients (e.g., January, November).

3. Design Segment-Specific Campaigns

- **Cluster 0**:
  - Run awareness campaigns targeting untapped potential with no prior campaign outcomes (`poutcome_nan`).
  - Use personalized, longer call durations to engage effectively.
- **Cluster 1**:
  - Re-engage clients with past failed outcomes using tailored incentives or improved messaging.
  - Optimize contact strategies in May to maximize returns.
- **Cluster 2**:
  - Emphasize affordability in product offerings for blue-collar workers with lower balances.
  - Focus on short, efficient call interactions to maintain engagement.

4. Improve Data Collection

- **Reduce Missing Data**:
  - Address gaps in critical features like `contact_nan` and `education_nan` to improve model accuracy.
  - Enhance data capture processes during client interactions.
- **Track Outcomes**:
  - For clients with `poutcome_nan`, implement mechanisms to record campaign results for future analysis.

5. Monitor and Iterate

- **A/B Test Campaigns**:
  - Test different strategies across clusters to identify the most effective approaches.
  - Evaluate campaign success using conversion rates and compare against logistic regression predictions.
- **Evaluate ROI**:
  - Regularly assess the return on investment for cluster-specific campaigns.
- **Update Models**:
  - Retrain clustering and logistic regression models periodically with new data to capture evolving customer behaviors.