BTA 350 ANALYTICS PROJECT

ANALYSIS: TERM DEPOSIT MARKETING CAMPAIGN

KRISTA GRAHAM

974886071

DECEMBER 4, 2022

TABLE OF CONTENTS

INTRODUCTION

The dataset that I selected from the Bank Marketing Data was:

bank-additiona-full.csv with all examples and 20 inputs, ordered by date (older version of this dataset with less inputs).

Task	Completion Due Date
Section 1: Introduction	October 2, 2022
Section 2: Context	October 9, 2022
Section 3: Defining the problem	October 16, 2022
Section 4: Data	October 23, 2022
Section 5: Analysing the Data (submit E.C.)	October 30, 2022
Section 6: Evaluation	November 6, 2022
Section 7: Interpretation	November 13, 2022
Section 8: Report	November 20, 2022
Section 9: Recommendations	November 27, 2022
Final Presentation + Analysis	December 4, 2022

Citation:

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

CONTEXT

The dataset for this project is from direct marketing campaigns of a banking institution in Portugal, and was collected via telephone calls to existing clients between the years of 2008 and 2013. The goal of collecting such data was to determine how successful telemarketing calls were for clients subscribing to long-term bank deposits. The data includes socio-economic, bank client, and product attributes. The approach I will be taking to this analytics project is a 7-step structured approach. To avoid any mistakes and ensure I have a quality outcome, I will complete all seven steps in their correct order as follows:

1) Define the problem
2) Get the right data
3) Analyze the data
4) Evaluate the results
5) Interpret the results
6) Communicate the results
7) Decide

DEFINING THE PROBLEM

1) **Problem definition:** The bank needs to focus their marketing efforts on clients who are at a higher likelihood of purchasing a term deposit.

2) **Question definition and link with the problem:** Out of the bank's existing customers, which demographics are most likely to sign up for a term deposit?

3) **Explanation of the business relevance of the problem:** By determining what types of clients are more likely to sign up for these deposits, the bank can adjust their marketing efforts to the most likely group(s), saving time and money.

DATA

The dataset is comprised of direct marketing campaign data of a Portuguese banking institution, specifically structured data which is easiest for me to analyse. The data was collected using a telemarketing system and was accessed from the UCI Machine Learning Repository. After relabelling the columns and cleaning things up, I have the following attributes:

Attribute	Type of Data & Description
Age	Numeric
Job	Categorical – type of job
Marital Status	Categorical
Education	Categorical – highest education level
Default	Categorical – does member have credit in default?
Housing	Categorical – has house loan?
Loan	Categorical – has personal loan?
Contact	Categorical – contact communication type
Month	Categorical – last contact month of year
Day_of_week	Categorical – what day of the week was the client last contacted?
Duration	Numeric – last contact duration of call in seconds
Campaign	Numeric – number of touches to client during this campaign
Previous	Numeric – number of touches to client prior to this campaign
pOutcome	Categorical – outcome of prior marketing campaign
Cons.price.idx	Numeric – consumer price index
Cons.conf.idx	Numeric – consumer confidence index
Subscribed	Binary Y/N – has client subscribed to term deposit?

To ensure that my data is appropriate for the above question, I began exploring the data by checking for errors, unknowns (null values) and plotting the different variables into charts to help me visualize. I created several PivotTables, conducting univariable analysis to familiarize myself with the data. I've provided a few tables and charts below.

Variable - Job Type:

% of Job	Column Labels		
Row Labels	no	yes	Grand Total
admin.	87.03%	12.97%	100.00%
blue-collar	93.11%	6.89%	100.00%
entrepreneur	91.48%	8.52%	100.00%
housemaid	90.00%	10.00%	100.00%
management	88.78%	11.22%	100.00%
retired	74.77%	25.23%	100.00%
self-employed	89.51%	10.49%	100.00%
services	91.86%	8.14%	100.00%
student	68.57%	31.43%	100.00%
technician	89.17%	10.83%	100.00%
unemployed	85.80%	14.20%	100.00%
unknown	88.79%	11.21%	100.00%
Grand Total	**88.73%**	**11.27%**	**100.00%**

Figure 1. PivotTable showing percentage of clients by job type as percent of row total

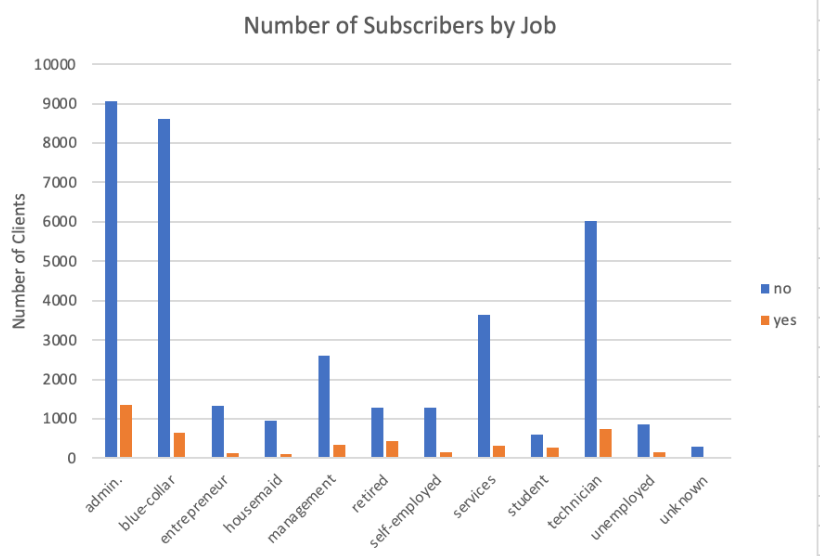


Figure 2. Bar chart showing the number (count) of clients by job type

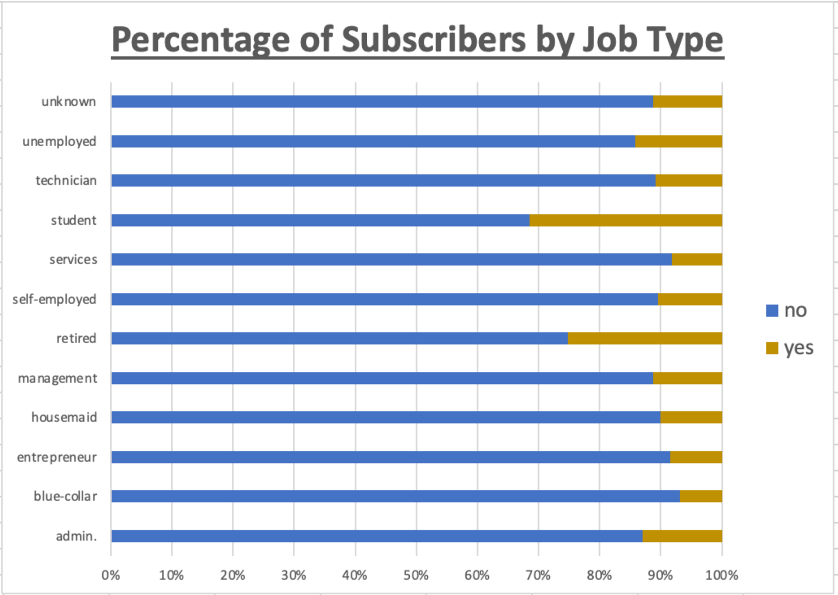


Figure 3. 100% stacked column chart showing percentage of clients by job type

<u>Variable – Contact Month:</u>

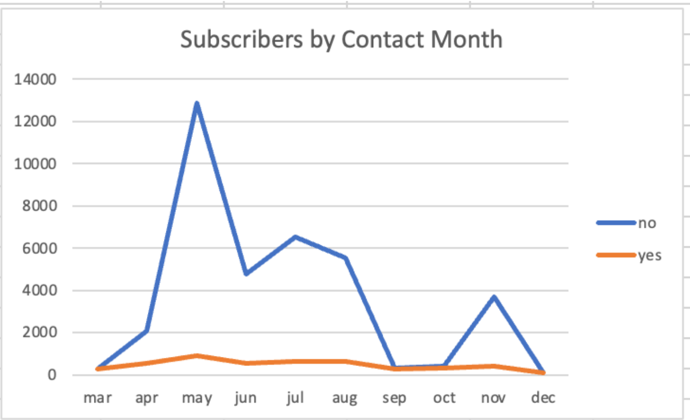


Figure 4. Line graph showing the number of clients by the month they were contact in

In figure 4, this plot shows a huge spike in “no” responses in May but likely due to it being the largest contact month. The “yes” line stays mostly unchanged regardless of the number of touches during the month.

<u>Variable - Education:</u>

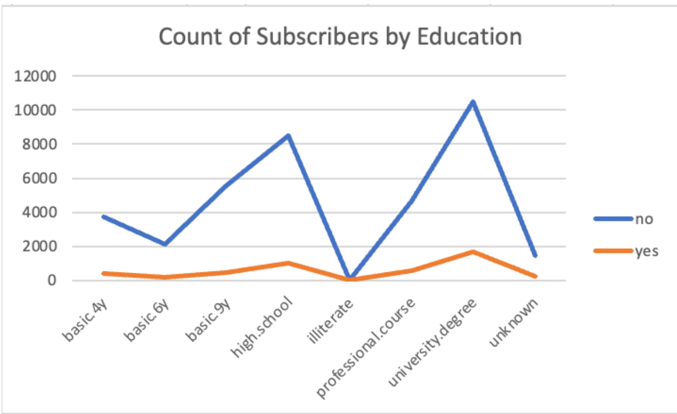


Figure 5. Count of clients by education level

I do see a few points of interest between the number of subscribers based on education as shown in Figure 5. There’s a noticeable spike in subscribership in those who have university degrees and high school educations. There’s a noticeable dip in subscribership in clients who are illiterate.

<u>Variable – Age:</u>

To evaluate client ages, I used Excel to group them into 10-year age brackets (<20, 20-29, 30-39, etc.) and created a PivotTable to see which percentage of each age group subscribed. Figure 7 shows a 100% stacked bar chart showing the PivotTable results from figure 6.

% of Age	Column Labels		
Row Labels	no	yes	Grand Total
<20	54.67%	45.33%	100.00%
20-29	84.13%	15.87%	100.00%
30-39	89.87%	10.13%	100.00%
40-49	92.08%	7.92%	100.00%
50-59	89.84%	10.16%	100.00%
60-69	65.33%	34.67%	100.00%
70-79	54.86%	45.14%	100.00%
80-89	48.57%	51.43%	100.00%
90-99	50.00%	50.00%	100.00%
Grand Total	**88.73%**	**11.27%**	**100.00%**

Figure 6: PivotTable showing percentage of clients by age as percent of row total

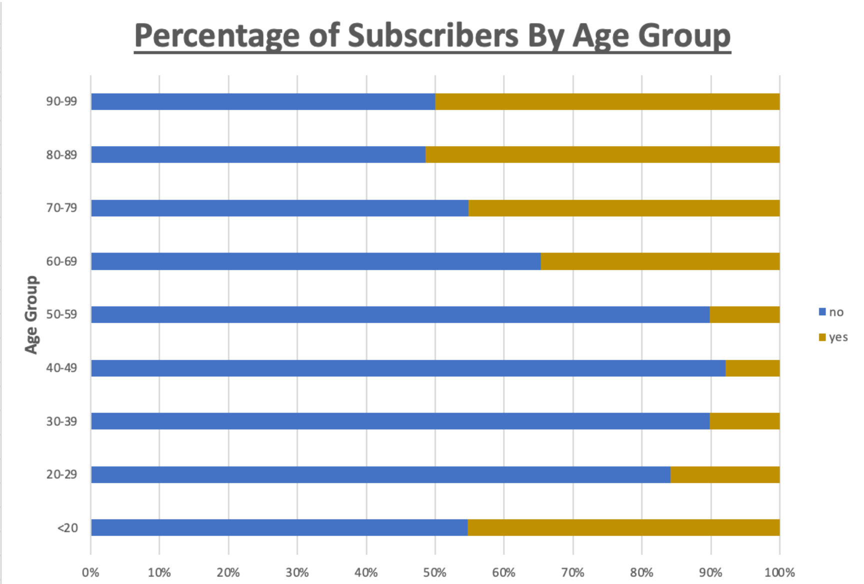


Figure 7. 100% stacked bar chart showing the percentage of clients based on age group

ANALYSIS

The outcome variable is whether a client subscribes to a term deposit or not ("yes" or "no"), with "yes" being the favourable outcome. By analysing what we know about the clients (age, job type, education level, etc.), we can use this information to see if certain demographics have a higher likelihood of purchasing a term deposit. During my analysis, I created two models that could be used by the bank to predict who will sign up for a term deposit.

Model 1:

According to the dataset, clients were contacted between 1 and 56 times during the marketing campaign. Of those 56 touches, 100% of subscribers did so within the first 23 touches and 67.69% said yes within the first 3 calls. There was a noticeable decrease in subscribership across all age groups after the first four touches. 96.92% of the total population falls between the ages of 20-59, and has a subscribership rate of 60.3% within the first three touches.

Formula: *If we contact clients between the ages of 20 and 59 three times per marketing campaign, we can expect a 60.30% subscription rate.*

Sum of Campaign	Column Labels									
Row Labels	<20	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	Grand Total
1	0.21%	4.64%	8.93%	4.28%	3.42%	1.33%	0.88%	0.43%	0.03%	24.16%
2	0.17%	4.94%	9.01%	4.68%	3.95%	1.53%	0.71%	0.42%	0.02%	25.44%
3	0.13%	3.50%	7.12%	3.03%	2.80%	0.88%	0.38%	0.22%	0.03%	18.09%
4	0.04%	1.97%	3.95%	2.10%	1.47%	0.59%	0.34%	0.00%	0.00%	10.46%
5	0.05%	1.37%	2.36%	1.21%	0.95%	0.21%	0.11%	0.05%	0.00%	6.30%
6	0.00%	0.63%	2.02%	0.95%	0.82%	0.13%	0.06%	0.13%	0.00%	4.73%
7	0.00%	0.74%	0.74%	0.15%	0.88%	0.07%	0.15%	0.07%	0.00%	2.79%
8	0.00%	0.08%	0.50%	0.34%	0.34%	0.08%	0.08%	0.00%	0.00%	1.43%
9	0.00%	0.28%	0.57%	0.09%	0.66%	0.00%	0.00%	0.00%	0.00%	1.61%
10	0.00%	0.11%	0.53%	0.42%	0.21%	0.00%	0.00%	0.00%	0.00%	1.26%
11	0.00%	0.00%	0.81%	0.46%	0.00%	0.12%	0.00%	0.00%	0.00%	1.39%
12	0.00%	0.00%	0.13%	0.25%	0.00%	0.00%	0.00%	0.00%	0.00%	0.38%
13	0.00%	0.14%	0.27%	0.14%	0.00%	0.00%	0.00%	0.00%	0.00%	0.55%
14	0.00%	0.00%	0.00%	0.15%	0.00%	0.00%	0.00%	0.00%	0.00%	0.15%
15	0.00%	0.16%	0.00%	0.00%	0.16%	0.00%	0.00%	0.00%	0.00%	0.32%
17	0.00%	0.00%	0.18%	0.18%	0.36%	0.00%	0.00%	0.00%	0.00%	0.71%
23	0.00%	0.00%	0.24%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.24%
Grand Total	**0.60%**	**18.55%**	**37.35%**	**18.41%**	**16.02%**	**4.95%**	**2.71%**	**1.32%**	**0.08%**	**100.00%**

Subscribed
no
yes

Figure 8. PivotChart showing percentage of clients by age group and number of touches
Note: Slicer used to filter only those who subscribed to term deposit

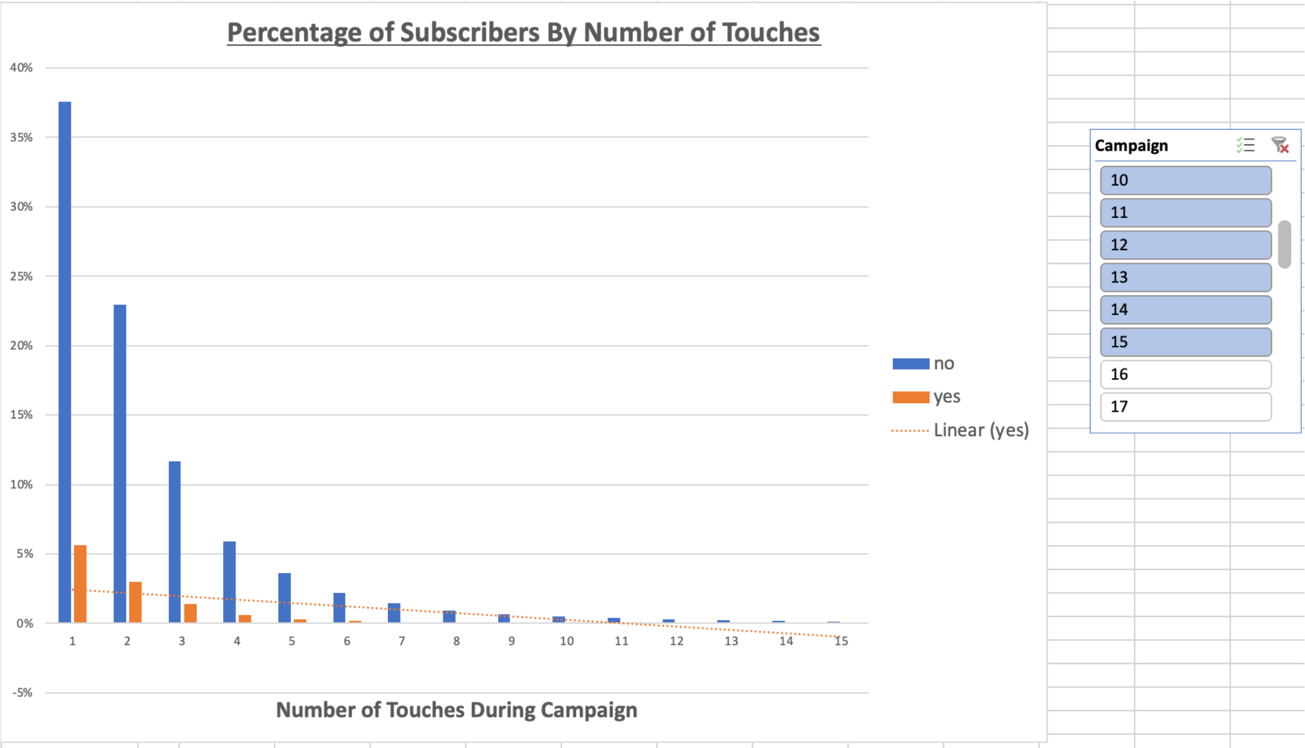


Figure 9. Bar Chart with trendline showing percentage of clients by number of touches

Model 2:

According to the data, the three top subscribing professions were admins, technicians, and blue-collar workers. Combined they made up 58.62% of those that subscribed. The highest subscribing age group amongst those three job categories was ages 30-39 (26.08% combined).

Formula: *If we focus our marketing efforts targeting admin, blue-collar and technician workers between the ages of 30 and 39, we can expect a 26.08% subscription rate.*

Count of Job	Column Labels									
Row Labels	<20	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	Grand Total
admin.	0.00%	6.34%	12.37%	5.63%	3.86%	0.86%	0.09%	0.00%	0.00%	29.14%
blue-collar	0.00%	2.03%	5.65%	3.51%	2.39%	0.15%	0.02%	0.00%	0.00%	13.75%
entrepreneur	0.00%	0.15%	1.08%	0.82%	0.52%	0.11%	0.00%	0.00%	0.00%	2.67%
housemaid	0.00%	0.11%	0.39%	0.47%	0.65%	0.41%	0.09%	0.17%	0.00%	2.28%
management	0.00%	0.58%	2.74%	1.64%	1.64%	0.39%	0.06%	0.02%	0.00%	7.07%
retired	0.00%	0.00%	0.02%	0.09%	2.05%	3.02%	2.76%	1.31%	0.11%	9.35%
self-employed	0.00%	0.58%	1.36%	0.82%	0.41%	0.04%	0.00%	0.00%	0.00%	3.21%
services	0.00%	1.75%	2.97%	1.55%	0.69%	0.00%	0.00%	0.00%	0.00%	6.96%
student	0.73%	4.18%	0.97%	0.04%	0.00%	0.00%	0.00%	0.00%	0.00%	5.93%
technician	0.00%	2.63%	8.06%	2.61%	2.11%	0.30%	0.02%	0.00%	0.00%	15.73%
unemployed	0.00%	0.60%	1.25%	0.67%	0.52%	0.06%	0.00%	0.00%	0.00%	3.10%
unknown	0.00%	0.19%	0.11%	0.13%	0.19%	0.06%	0.06%	0.04%	0.00%	0.80%
Grand Total	**0.73%**	**19.14%**	**36.96%**	**17.97%**	**15.02%**	**5.41%**	**3.10%**	**1.55%**	**0.11%**	**100.00%**

Subscribed
no
yes

Job
admin.
blue-collar
entrepreneur
housemaid
management
retired
self-employed
services

Figure 10. PivotChart showing percentage of clients by age group and job type
Note: Slicer used to filter only those who subscribed to term deposit

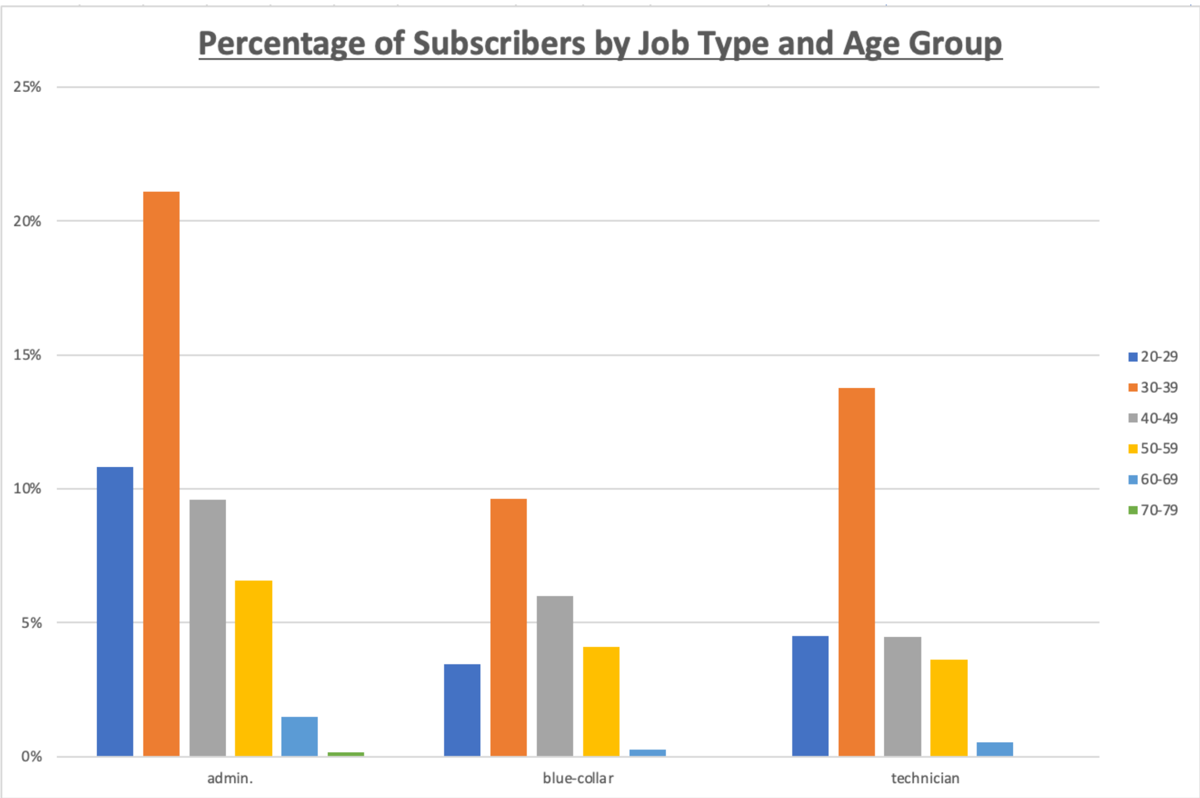


Figure 11. Bar Chart showing percentage of subscribers by age group and job type

EVALUATION

Confounding:

Model One: The marketing campaign was restricted to strictly telephone (mobile and landline) touches. It's possible that the number of contacts in past campaigns affected whether they subscribed or not. For example, if they were recently contacted in a prior campaign that would affect the number of times the client was ever contacted.

Another alternative, is what time of day was the client contacted? Since the touches are done via phone, perhaps the client was more available due to work schedule.

Model Two: It's possible that clients between the ages of 30-39 are either married and/or have children, which could be a factor in their decision.

Overfitting:

The bank targeted 41,189 clients across 11 professions for this marketing campaign. While the total number of clients is unknown, that seems to be a considerable sample size for this campaign. An area of opportunity for future campaigns is to expand or include other forms of marketing (mailers, digital contact such as text or email, and in-person promotions at specific bank branches).

Causality:

To ensure that my claims are not based simply on correlation, its important to consider the clients' characteristics are similar enough across all clients in the claim. For example, model one has a single method of contact (phone calls) versus a variety of methods. By adding in email, fax, in-person marketing or mailers, it may require a different number of touches per client to get them to subscribe. I am confident that the number of touches by telephone is causally related and not just a correlation. For model two, it is possible that there could be additional factors affecting whether these job categories have a higher likelihood of subscribing. Income could be a varying factor that might affect the clients' willingness to subscribe.

Significance:

Model One: It's clear that the financial institution has invested a significant amount of time and money into this marketing campaign. The data shows that a client can be contacted up to 53 times. Model one suggests decreasing the number of touches to just three per client within the ages of 20 and 59. This will have a significant financial impact by removing ineffective marketing attempts.

Model Two: 37% of clients fell into the age group of 30-39, making it the largest group. By marketing to individuals in the top three job types, it may be too small of a sample size for us to state whether this model would be significant enough to implement.

Effect size:

The dataset had data for a total of 41,189 clients. For both models, I wanted to ensure that I wasn't using too small of a sample size to avoid effect size impacts. For example, the data shows that 50% of clients between the ages of 90-99 subscribed, which seems like a large percentage except that that age group only made up 0.02% of all subscribers. By using the ages 20-59 for model one which includes nearly 97% of clients, it will ensure that's slight variations in repeated tests will not have a large impact on the results.

For model two, I did use a smaller age group but a larger pool of job types to minimize effect size issues, utilizing 29.22% of the population. I do not think that model two would be as resilient to effect size issues than the first model.

Final Thoughts:

Overall, I think that model one is a solid option and I'm quite satisfied with the results. Initially I thought that model two was the better option, however it has issues with effect size and significance. I truly think that model one is the best option of the two because it has a larger sample size with fewer confounding relationships.

INTERPRETATION

Now that my analysis and evaluation of results is complete, I'm ready to interpret the results as it pertains to the original question, which was: "Out of the bank's existing customers, which factors affect who will sign up for a term deposit?". As I mentioned previously using model one, 60.3% of clients who subscribed were between the ages of 20 and 59 and did so within the first three campaign touches. This sample size also made up 50.21% of the total population. The directionality of number of touches and likelihood to subscribe is has a negative correlation, because the percentage of those that subscribe decreases as the number of touches increases.

Since the dataset had data for a total of 41,189 clients, I wanted to ensure that I wasn't using too small of a sample size to avoid issues with magnitude. By using the ages 20-59 for model one which includes nearly 97% of clients, it will ensure that's slight variations in repeated tests will not have a large impact on the results. It also removes what I would consider outliers, which are ages above 90 and below 20 since there were only a few clients in those age groups.

RECOMMENDATIONS

It is my recommendation to the client that marketing efforts be focused on clients who are between the ages of 20 and 59, with a maximum number of three touches per client. This shift in marketing efforts will save time and money because marketing will be focused on the demographic most likely to subscribe to a term deposit.

To maintain transparency with clients and to ensure privacy risks are minimized, I recommend that they are provided a copy of marketing policies with the option of opting out if they so choose.

Approach	Details	Data	Benefit	Risk
Model 1 (preferred)	If we contact clients ages 20-59 3x per campaign, we can expect a 60.30% subscription rate.	Age groups between 20 and 59 Number of phone calls during current campaign	Targets over 50% of total client population, focusing on those most likely to subscribe.	Does not factor in other methods of marketing (email, mailers, etc.)
Model 2	If target admin, blue-collar and technician workers between the ages of 30-39, we can expect a 26.08% subscription rate.	Single age group between 30 and 39 <u>Three Job Types:</u> 1) Admin 2) Blue-Collar 3) Technicians	Targets top three converting job types Narrows the age target group	Less of a sample size (significance) Not as resilient to effect size issues

EXECUTIVE SUMMARY

Overview

The bank conducted a telemarketing campaign between 2008 - 2010 and provided the results of whether a client subscribed to a term deposit (outcome variable). After analysing the client client demographics, I have provided my recommendation for a more strategic and cost-effective marketing campaign.

Problem Statement

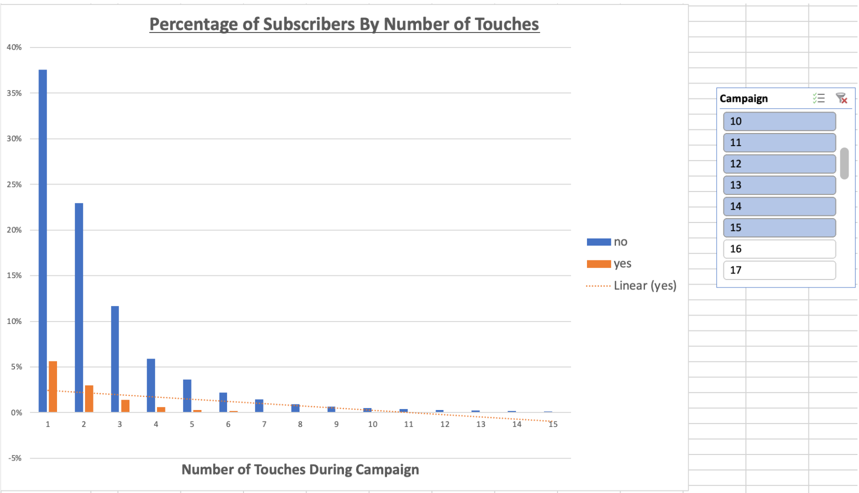
The bank needs to focus their marketing efforts on clients who are at a higher likelihood of purchasing a term deposit.

Data and Analysis

To ensure that the data can answer the above question, I began exploring the data by checking for errors, unknowns (null values) and plotting the different variables into charts to help me visualize any issues or outliers. From there, I created two data models that could be used to market more effectively to clients. I then tested the two models to check for any issues or errors.

Proposed Solution

It is my recommendation that the bank targets clients who are between the ages of 20-59, with a maximum number of three marketing contacts (touches) per client, using Model One. As you can see in the graph below, 50% of subscribers do so after the first call, and drops to 5% after the third call.



Value

This shift in marketing efforts will save time and money because marketing will be focused on the demographic most likely to subscribe to a term deposit.

PRESENTATION

[Please review a recording of my presentation here.]