MKTG 562: Group Assignment 1

Aishwary Jadhav, Jaivardhan Chauhan, Maitreyi Ekbote, Shreyansh Bhatia

Part I) Prediction

a)

First, we created the Frequency using NumStorePurchases and NumWebPurchases columns, as given in the question. Similarly, Monetary column was created using six categories of MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, and MntGoldProds columns. We also calculated the RFM score and attached it to the dataframe.

Given variables are cost per contact and profit margin. We can calculate the profit per customer by subtracting these values, which is 8. Total profit can be calculated using the number of customers who have responded yes and the profit per customer.

If we target everyone in the sample of 2240 customers, the total profit is <u>-4048$</u>.

<u>Insights</u>

Targeting all 2,240 customers in the dataset without segmentation leads to a total loss of -$4,048. This occurs because while some customers respond positively, the overall response rate is too low to cover the fixed marketing costs of reaching out to everyone. By contacting all customers, the campaign includes a significant number of low-probability responders, leading to unnecessary expenses that outweigh revenue generation.

<u>Why Does Targeting Everyone Lead to a Loss?</u>

The negative profit of -$4,048 shows the limitations of untargeted marketing. Since each customer contact incurs a fixed cost of $3 per offer, but only a fraction of customers respond positively, the total expenditure outweighs revenue. The company spends resources on individuals who are unlikely to engage, reducing the overall return on investment (ROI). Without data-driven selection, the campaign fails to focus on high-probability customers, leading to significant financial losses.

R Code:

```
df <- df %>% mutate(Frequency = NumWebPurchases + NumStorePurchases)
df <- df %>% mutate(Monetary = MntWines + MntFruits + MntMeatProducts +
MntFishProducts + MntSweetProducts + MntGoldProds)

R_temp <- ntile(df$Recency, 3)
df$R_score <- factor(R_temp, levels = 1:3)

F_temp <- 4 - ntile(df$Frequency, 3)
```

```
df$F_score <- factor(F_temp, levels = 1:3)

M_temp <- 4 - ntile(df$Monetary, 3)
df$M_score <- factor(M_temp, levels = 1:3)

df <- df %>%
  mutate(
    RFM_score = 100 * as.numeric(as.character(R_score)) +
      10 * as.numeric(as.character(F_score)) +
      as.numeric(as.character(M_score))
  )

cost_per_contact = 3
profit_margin = 11
profit_per_customer <- profit_margin - cost_per_contact

positive_responses <- sum(df$Response)
total_cost <- cost_per_contact * nrow(df)
total_profit <- (positive_responses * profit_per_customer) - total_cost
```

---

b)

If the company uses RFM, the profit will be -31$.
Break-Even Response Rate: 27.27%
Number of Customers Targeted in RFM Model: 333

Insights

Using RFM-based targeting, where only high-value customers (those exceeding the break-even response rate of 27.27%) are selected, results in a significant reduction in losses, bringing the total profit to - $31, compared to - $4,048 when targeting all customers. This improvement highlights the effectiveness of segmentation, as the campaign now avoids contacting low-value customers who are unlikely to respond. Although the campaign is still operating at a slight loss, it is nearly breaking even, suggesting that RFM helps optimize marketing spend.

Why Does RFM Improve Profitability Over Targeting Everyone?

The improvement from -$4,048 to -$31 occurs because RFM-based targeting prioritizes high-value customers that is, those who purchase frequently, spend more, and have recently engaged with the company. Instead of wasting marketing resources on low-response customers, RFM ensures that only customers with a higher likelihood of responding are contacted, leading to a higher average response rate and lower overall marketing costs. However, RFM segmentation has limitations, as it relies on fixed cutoffs rather than calculating each customer's exact probability of responding. To further improve profitability, a regression-based approach could refine targeting even more effectively, selecting only the most promising customers rather than grouping them into predefined RFM segments.

R Code:

```
break_even_rate <- cost_per_contact / profit_margin
print(paste("Break-Even Response Rate:", round(break_even_rate * 100, 2), "%"))

df_rfm <- df %>%
  group_by(RFM_score) %>%
  mutate(response_prob = mean(Response))

target_customers_rfm <- df_rfm %>% filter(response_prob > break_even_rate)

respondents_rfm <- sum(target_customers_rfm$Response)
profit_rfm  <-  (profit_per_customer  *  respondents_rfm)  -  (cost_per_contact  *
nrow(target_customers_rfm))
avg_response_rate_rfm <- mean(target_customers_rfm$Response)
```

-

c)

The regression results are below:

```
Call:
lm(formula = Response ~ Recency + Frequency + Monetary, data = df_rfm)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.60719 -0.18043 -0.09484  0.00004  1.02315 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.473e-01  1.955e-02  12.650  < 2e-16 ***
Recency     -2.548e-03  2.434e-04 -10.471  < 2e-16 ***
Frequency   -1.121e-02  1.874e-03  -5.982 2.56e-09 ***
Monetary     2.272e-04  1.628e-05  13.951  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3333 on 2236 degrees of freedom
Multiple R-squared:  0.1259,	Adjusted R-squared:  0.1248 
F-statistic: 107.4 on 3 and 2236 DF,  p-value: < 2.2e-16
```

The coefficient for Recency is negative (-0.002548, $p < 0.001$), meaning that as Recency increases (i.e., the last purchase was made a long time ago), the likelihood of responding to the offer decreases. This confirms that recent customers are more engaged and more likely to respond to new promotions, while those who have not purchased in a long time are less likely to return. This insight suggests that targeting customers who have purchased recently is more effective, as they demonstrate a higher level of brand engagement.

The coefficient for Frequency is also negative (-0.01121, $p < 0.001$), suggesting that customers who purchase more frequently are actually less likely to respond to the offer. One possible explanation is that frequent shoppers may already have a sufficient supply of the product or are already engaged enough that additional offers do not significantly increase their likelihood of responding. Another possibility is that frequent buyers expect better loyalty rewards rather than standard promotional offers, indicating a potential need for differentiated marketing strategies for high-frequency shoppers.

On the other hand, Monetary value has a strong positive coefficient (0.0002272, $p < 0.001$), indicating that customers who spend more are more likely to respond to the offer. This suggests that high-value customers, even if they do not shop frequently, should be a primary focus for marketing campaigns. These customers are more receptive to promotions and offers, making them a key segment for personalized engagement strategies aimed at maximizing revenue.

R Code:

```
model <- lm(Response ~ Recency + Frequency + Monetary, data = df_rfm)
summary(model)
```

---

d)

The average response rate under this regression model is <u>0.3964</u>

The average response rate in Part b) under the RFM model is <u>0.3634</u>

This shows that the regression model performs better than the RFM model in predicting customer responses.

<u>Insights</u>

The results show a clear advantage of regression-based targeting over RFM segmentation. The average response rate under the regression model is 39.64%, compared to 36.34% under RFM, demonstrating that regression more accurately identifies high-response customers. Unlike RFM, which assigns customers to broad categories, regression calculates an individualized probability of response, ensuring that only the most likely responders are targeted. This reduces wasted marketing efforts on low-response customers, improving efficiency. Additionally, the regression model confirms that Monetary value has the strongest positive impact, reinforcing that high spenders are the most receptive to promotions, which is an insight that RFM may not fully capture.

Why Does Regression Improve Response Rates?

Regression improves response rates by selecting customers based on precise probability estimates rather than broad segmentation rules. While RFM assumes that customers within the same category behave similarly, regression accounts for continuous differences in behavior, leading to better prioritization of high-value customers. As a result, the regression model yields a higher response rate, making it a more effective and profitable targeting approach. This confirms that predictive modeling should be the preferred method for customer selection, with RFM serving as a useful but less refined segmentation tool.

R Code:

```
num_rfm_targeted <- nrow(target_customers_rfm)
df_rfm$Predicted_Response <- predict(model, type = "response")

df_regression <- df_rfm %>% arrange(desc(Predicted_Response))
regression_targeted_customers <- df_regression %>% head(num_rfm_targeted)

avg_response_rate_regression <- mean(regression_targeted_customers$Response)
```

——
–

e)

The profit under this regression model is $57.
The profit in Part b) under the RFM model is -$31

Insights

Using regression-based targeting, the campaign achieves a profit of $57, a clear improvement over the -$31 loss under RFM-based targeting. This shift from a loss to profitability highlights the advantage of predictive modeling over traditional segmentation. Unlike RFM, which groups customers into fixed categories, regression assigns an individual probability of responding, ensuring that only high-likelihood responders are targeted. This reduces marketing costs by avoiding low-response customers and maximizes revenue by prioritizing high-value responders.

Why Does Regression Improve Profitability Over RFM?

Regression-based targeting outperforms RFM because it uses data-driven predictions rather than static segmentation rules. While RFM segments customers into broad categories, regression ranks them based on their exact likelihood of responding, leading to a more refined and cost-effective selection process. The shift from -$31 to +$57 proves that a probability-based approach increases profitability, confirming that predictive modeling should be the preferred method for future campaigns, with RFM serving as an initial filtering tool before applying regression-based ranking.

R Code:

```
regression_positive_responses <- sum(regression_targeted_customers$Response)
regression_total_cost <- cost_per_contact * nrow(regression_targeted_customers)
regression_total_profit <- (regression_positive_responses * profit_per_customer) -
regression_total_cost
```

——
-

f)

AIC for Original Model: 1440.6972

AIC for Extended Model: 1445.8909

```
Call:
lm(formula = Response ~ Recency_Sq + Frequency_Sq + Monetary_Sq +
    R_F + R_M + F_M, data = df_reg_new)

Residuals:
     Min       1Q   Median       3Q      Max
-0.89068 -0.15038 -0.10834 -0.02335  1.03356

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.608e-01  1.424e-02  11.291  < 2e-16 ***
Recency_Sq   -1.825e-05  4.484e-06  -4.071 4.85e-05 ***
Frequency_Sq -3.190e-04  2.056e-04  -1.552  0.12082
Monetary_Sq   1.301e-07  2.637e-08   4.934 8.65e-07 ***
R_F           4.234e-05  5.463e-05   0.775  0.43834
R_M          -1.382e-06  5.099e-07  -2.711  0.00676 **
F_M           1.675e-06  4.416e-06   0.379  0.70458
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3335 on 2233 degrees of freedom
Multiple R-squared:  0.1263,    Adjusted R-squared:  0.1239
F-statistic: 53.78 on 6 and 2233 DF,  p-value: < 2.2e-16
```

The extended regression model, which includes interaction terms and squared variables, does not improve model fit, as seen in the AIC increase from 1440.70 to 1445.89. This suggests that adding nonlinear and interaction effects does not enhance predictive accuracy. While Recency^2, Frequency^2, and Monetary^2 are statistically significant ($p < 0.01$), most interaction terms (R × F, R × M, F × M) remain insignificant, indicating that Recency, Frequency, and Monetary influence customer response independently rather than interactively.

The results align with expected RFM behavior, where Recency remains a strong negative predictor ($p < 0.001$), confirming that recent customers are more likely to respond. However, the limited impact of interaction terms suggests that spending and purchase frequency contribute individually rather than amplifying each other's effects. Since the AIC increased and interaction terms were mostly insignificant, the simpler model remains more interpretable and effective for customer targeting.

R Code:

```
df_reg_new <- df_regression %>%
  mutate(
    Recency_Sq = I(Recency^2),
    Frequency_Sq = I(Frequency^2),
    Monetary_Sq = I(Monetary^2),
    R_F = Recency * Frequency,
    R_M = Recency * Monetary,
    F_M = Frequency * Monetary
  )

model_extended <- lm(Response ~ Recency_Sq + Frequency_Sq + Monetary_Sq + R_F + R_M
+ F_M, data = df_reg_new)
aic_extended <- AIC(model_extended)
summary(model_extended)
```

-

g)

Optimal Threshold (δ): 0.33
Total Profit at Optimal δ: 230
Number of Customers Targeted at Optimal δ: 198
Customers Targeted in Part (d) (RFM-Based): 333

Insights

By using a purely regression-based targeting approach, the model achieves an optimal profit of $230, an improvement over the -$31 loss from RFM-based targeting in Part (b). This confirms that data-driven, probability-based selection is more effective than predefined RFM segmentation rules. The optimal threshold (δ) of 0.33 ensures that only customers with at least a 33% predicted response probability are targeted. This results in 198 customers being selected, compared to 333 in RFM-based targeting (Part d), showing that regression-based targeting is more efficient, as it focuses on fewer but higher-response customers, increasing overall profitability.

Why Does Regression-Only Targeting Improve Profitability?

Unlike RFM-based selection, which preselects a fixed number of customers, regression-based targeting dynamically determines the optimal number based on predicted response likelihood.

In Part (d), RFM targeted 333 customers, including some lower-response individuals, leading to unnecessary marketing costs. By selecting only 198 highly probable responders, the regression model minimizes wasteful spending while improving profitability ($230 vs. -$31). This confirms that regression-based targeting should be the preferred approach, as it optimizes customer selection and marketing ROI, outperforming RFM segmentation-based strategies.

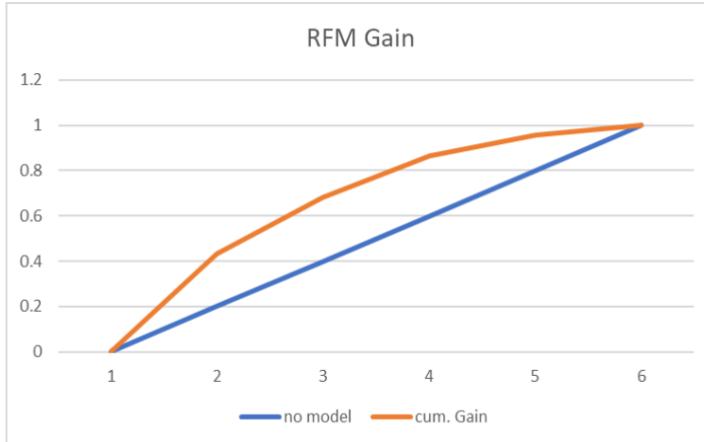
R Code:

```
df_reg_new$Predicted_Response <- predict(model_extended, type = "response")

calculate_profit <- function(threshold) {
  targeted_customers <- df_reg_new %>% filter(Predicted_Response >= threshold)
  num_targeted <- nrow(targeted_customers)

  if (num_targeted == 0) {
    return(-Inf)
  }

  positive_responses <- sum(targeted_customers$Response)
  total_cost <- cost_per_contact * num_targeted
  total_profit <- (positive_responses * profit_per_customer) - total_cost

  return(total_profit)
}

threshold_values <- seq(0, 1, by = 0.01)
profit_results <- sapply(threshold_values, calculate_profit)
optimal_threshold <- threshold_values[which.max(profit_results)]
optimal_profit <- max(profit_results)


optimal_targeted_customers <- df_reg_new %>% filter(Predicted_Response >=
optimal_threshold)
num_optimal_targeted <- nrow(optimal_targeted_customers)
```

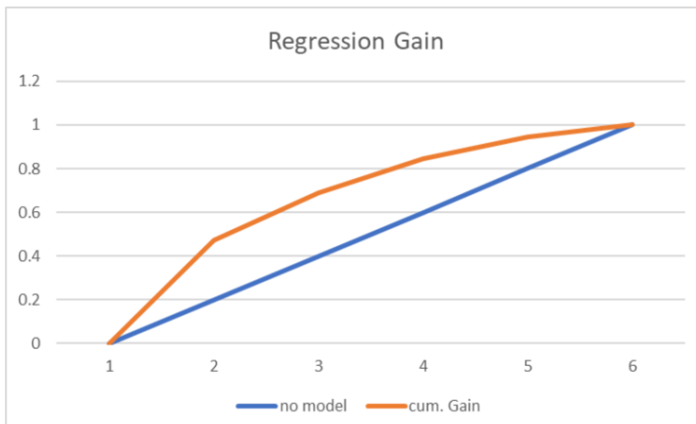
-

h)

	rfm_score_tiles	count	buyers
1	1	448	145
2	2	448	82
3	3	448	62
4	4	448	31
5	5	448	14



	regression_score_tiles	count	buyers
1	1	448	158
2	2	448	72
3	3	448	52
4	4	448	34
5	5	448	18



Insights

The regression-based gain graph outperforms RFM by more accurately ranking high-response customers. The top quintile in regression captures more buyers (158 vs. 145 in RFM), and its steeper gain curve shows better efficiency in identifying responders earlier. This reduces wasted marketing efforts and improves engagement with high-value customers.

RFM segments customers into broad groups, leading to a sharper drop in buyers across quintiles. Regression assigns individualized probabilities, resulting in a smoother distribution and a consistently higher lift over a no-model scenario. This confirms that regression-based targeting is a more effective strategy for maximizing response rates and profitability.

R Code:

```
df_rfm_grouped <- df_rfm %>%
  group_by(RFM_score) %>%
  mutate(buyprob_iq = mean(Response))
```

```
df_rfm_grouped$rfm_score_tiles <- 6 - ntile(df_rfm_grouped$buyprob_iq, 5)
lift_gains_data <- df_rfm_grouped %>%
  group_by(rfm_score_tiles) %>%
  summarise(
    count = n(),
    buyers = sum(Response == 1),
  )
df_new_regression  <- df_reg_new %>% mutate(Predicted_Response)
df_new_regression$regression_score_tiles             <-             6              -
ntile(df_new_regression$Predicted_Response, 5)
lift_gains_data_regression <- df_new_regression %>%
  group_by(regression_score_tiles) %>%
  summarise(
    count = n(),
    buyers = sum(Response == 1)  # Actual buyers to analyze lift and gains
  )
```

---

i)

Adjusted R-Squared (earlier model): 0.1234

Adjusted R-Squared (new model): 0.2411

Optimal Threshold (δ) for New Model: 0.36

Total Profit at Optimal δ (New Model): 453

Number of Customers Targeted at Optimal δ (New Model): 233

```
Call:
lm(formula = Response ~ Recency + Frequency + Recency_Sq + Frequency_Sq +
    Monetary_Sq + R_F + R_M + F_M + Income + Prev_Response_Sum,
    data = df_best)

Residuals:
     Min       1Q   Median       3Q      Max
-1.04835 -0.15604 -0.05043 -0.00558  1.00865

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.234e-01  3.917e-02   5.703 1.34e-08 ***
Recency           -5.903e-03  9.998e-04  -5.904 4.10e-09 ***
Frequency          1.473e-02  6.248e-03   2.358 0.018460 *
Recency_Sq         3.076e-05  8.833e-06   3.483 0.000506 ***
Frequency_Sq      -1.005e-03  3.026e-04  -3.323 0.000905 ***
Monetary_Sq        7.397e-08  2.548e-08   2.904 0.003723 **
R_F                1.263e-04  5.837e-05   2.163 0.030633 *
R_M               -1.369e-06  4.863e-07  -2.816 0.004908 **
F_M                1.318e-06  4.256e-06   0.310 0.756782
Income            -8.797e-07  3.508e-07  -2.508 0.012224 *
Prev_Response_Sum  1.934e-01  1.120e-02  17.268  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3113 on 2205 degrees of freedom
Multiple R-squared:  0.2449,    Adjusted R-squared:  0.2415
F-statistic: 71.53 on 10 and 2205 DF,  p-value: < 2.2e-16
```

Insights

Incorporating Income and Previous Campaign Responses significantly improved model performance, with Adjusted R-squared rising from 0.1234 to 0.2411. Previous campaign responses (Prev_Response_Sum) had the strongest positive effect ($p < 0.001$), confirming that past engagement predicts future response. Income showed a negative effect ($p = 0.0122$), suggesting lower-income customers are more responsive, possibly due to price sensitivity.

With the new model, the optimal threshold (δ) is 0.36, targeting 233 customers, up from 198 in the previous regression model. This increases profit to $453, surpassing the $230 from prior regression targeting and the -$31 loss in RFM-based targeting. By leveraging demographics and past behavior, the model enhances targeting precision, reduces waste, and boosts profitability, proving that regression-based approaches should use all available data for optimal customer selection.

R Code:

```
df_best <- df_reg_new %>%
  ungroup() %>%
  mutate(
    Marital_Status = as.factor(Marital_Status),
    Prev_Response_Sum = rowSums(select(., AcceptedCmp1, AcceptedCmp2, AcceptedCmp3,
AcceptedCmp4, AcceptedCmp5), na.rm = TRUE)
  ) %>%
  filter(!is.na(Income))
```

```
model_best <- lm(Response ~ Recency + Frequency + Recency_Sq + Frequency_Sq +
Monetary_Sq +
                             R_F + R_M + F_M +
                             Income + Prev_Response_Sum,
                           data = df_best)
summary(model_best)

adjusted_r_squared_original <- 1 - (1 - summary(model_extended)$r.squared) *
  ((nrow(df_best) - 1) / (nrow(df_best) - length(coef(model_extended)) - 1))

adjusted_r_squared_new <- 1 - (1 - summary(model_best)$r.squared) *
  ((nrow(df_best) - 1) / (nrow(df_best) - length(coef(model_best)) - 1))


df_best$Predicted_Response_New <- predict(model_best, type = "response")

calculate_profit_new <- function(threshold) {
  targeted_customers <- df_best %>% filter(Predicted_Response_New >= threshold)
  num_targeted <- nrow(targeted_customers)

  if (num_targeted == 0) {
    return(-Inf)  # Avoid division issues
  }

  positive_responses <- sum(targeted_customers$Response)
  total_cost <- cost_per_contact * num_targeted
  total_profit <- (positive_responses * profit_per_customer) - total_cost


  return(total_profit)
}

threshold_values <- seq(0, 1, by = 0.01)
profit_results <- sapply(threshold_values, calculate_profit_new)

optimal_threshold_new <- threshold_values[which.max(profit_results)]
optimal_profit_new <- max(profit_results)

optimal_targeted_customers_new <- df_best %>% filter(Predicted_Response_New >=
optimal_threshold_new)
num_optimal_targeted_new <- nrow(optimal_targeted_customers_new)
```

-

Part II) Causality

a) The RCT analysis shows ads caused a statistically significant 0.77 percentage point increase in conversion probability.

Causal Impact Calculation:

Conversion Rates:
Treatment (ad): 2.55%
Control (psa): 1.79%

Absolute Impact = 0.0255 − 0.0179 = +0.77% (percentage points)

Relative Improvement = (0.0077/0.0179)*100 = +43.1%

t-statistic: 8.66
p-value: 2.2e-16
The extremely small p-value (<0.0001) indicates we can confidently reject the null hypothesis that ads have no effect on conversions.

R Code:

```r
# Load libraries
library(readr)

# Import data
data <- read_csv("marketin_AB.csv")

# Convert response to binary
data$converted <- as.integer(as.logical(data$converted))

# Calculate conversion rates
conversion_rates <- aggregate(converted ~ test.group, data, mean)
ad_rate <- conversion_rates[conversion_rates$test.group == "ad", "converted"]
psa_rate <- conversion_rates[conversion_rates$test.group == "psa", "converted"]

# Calculate impact metrics
absolute_impact <- ad_rate - psa_rate
relative_improvement <- (absolute_impact / psa_rate) * 100

# Perform t-test
t_test_result <- t.test(converted ~ test.group, data = data)

# Output results
print(conversion_rates)
cat("\nAbsolute Impact:", absolute_impact)
cat("\nRelative Improvement:", relative_improvement, "%")
print(t_test_result)
```

b) The analysis reveals significant variation in ad effectiveness across prior ad exposure levels, with the strongest impact in high-exposure customers and unexpected patterns in moderate-exposure groups.

Causal Impact by Ad Exposure Quintile

quintile	ad	psa	absolute_impact	ad_rate	psa_rate	p_value
<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0.00217	0.00142	0.000742	0.00217	0.00142	1.81e- 1
2	0.00372	0.00524	-0.00152	0.00372	0.00524	1.31e- 1
3	0.00733	0.00770	-0.000363	0.00733	0.00770	7.99e- 1
4	0.0166	0.0152	0.00144	0.0166	0.0152	4.31e- 1
5	0.0978	0.0606	0.0372	0.0978	0.0606	7.75e-25

Strongest Impact - Q5 (High Exposure):

a. +3.72% absolute impact (p<0.001)

b. It shows high-exposure customers are primed for conversion through repeated engagement
c. This can suggest cumulative ad effectiveness or targeting refinement

Weakest Impact - Q2 (Moderate-Low Exposure):
a. Negative impact of -0.15%
b. There is a potential ad fatigue or mismatch between the ad content and mid-funnel customers
c. This indicates need for frequency capping or creative refresh

Further analyzing, we can see that there is minimal/no incremental impact for the middle quintiles (Q3, Q4). This potentially highlights diminishing returns beyond the initial exposures.

From a marketing perspective, more budget should be allocated to high-frequency ad viewers to double-down on the high exposure customers. The strategy for the low exposure customers should be reassessed by investigating why moderate exposure correlates with reduced effectiveness. Also, for such customers, limit ads to avoid negative sentiment. Optimize mid-funnel by developing targeted campaigns for middle quintiles to break through plateau effects.

The results suggest non-linear response patterns where only the highest-exposure group shows economically meaningful positive impacts, highlighting the importance of exposure-level targeting in campaign optimization.

<u>R Code:</u>

```r
library(dplyr)
library(broom)

# Load and prepare data
data <- read_csv("marketin_AB.csv") %>%
  mutate(converted = as.numeric(as.logical(converted)),
         quintile = ntile(total.ads, 5))

# Calculate conversion rates and impacts
results <- data %>%
  group_by(quintile, test.group) %>%
  summarise(conversion_rate = mean(converted), .groups = "drop") %>%
  pivot_wider(names_from = test.group, values_from = conversion_rate) %>%
  mutate(absolute_impact = ad - psa)

# Perform t-tests
tests <- data %>%
  group_by(quintile) %>%
  group_modify(~ tidy(t.test(converted ~ test.group, data = .x))) %>%
  select(quintile, estimate1, estimate2, p.value)

# Combine results
final_results <- left_join(results, tests, by = "quintile") %>%
  rename(ad_rate = estimate1, psa_rate = estimate2, p_value = p.value)

print(final_results)
```

---

c) The causal impact of CCPA implementation is captured by the treatment:after coefficient ($\eta = 0.0742$).

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.3018616  0.0012097 249.534   <2e-16 ***
treatment       0.0007211  0.0017108   0.422    0.673
after           0.0020817  0.0017108   1.217    0.224
treatment:after 0.0741897  0.0024194  30.664   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1951 on 103996 degrees of freedom
Multiple R-squared:  0.02773,	Adjusted R-squared:  0.0277
F-statistic: 988.8 on 3 and 103996 DF,  p-value: < 2.2e-16
```

The interaction term treatment:after has a p-value < 0.001, indicating strong statistical significance. The coefficient of 0.0742 implies that California customers experienced a 7.42 percentage point increase in their online purchase ratio (relative to the control group) after CCPA implementation. The intercept (baseline online purchase ratio for control states pre-CCPA) is 0.302 (30.2%). A 7.42 percentage point increase represents a 24.6% relative lift from baseline (0.074 / 0.302 ≈ 0.246).

This result aligns with two plausible mechanisms - privacy-conscious customers in California may have increased online purchases due to enhanced trust in data practices under CCPA, and despite potential compliance costs, companies may have invested in privacy infrastructure to retain/grow customer engagement, offsetting operational burdens.

The positive, significant η supports the hypothesis that CCPA increased online purchase ratios in California relative to control states. While surprising given initial concerns about regulatory costs, the result is plausible if privacy protections boosted consumer confidence more than compliance costs deterred business investment.

<u>R Code:</u>

```r
# Load necessary libraries
library(tidyverse)

# Load and prepare data (replace with actual data path)
privacy_data <- read_csv("privacy.csv")

model <- lm(online_ratio ~ treatment * after,
              data = privacy_data)

# Coefficient Table
summary(model)
```

---

d) The analysis reveals distinct patterns in how CCPA's impact varies across demographic groups.

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.40804 -0.14430 -0.01489  0.12431  1.22216 

Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                3.099e-01  6.325e-03  49.001  < 2e-16 ***
treatment                                 -9.117e-04  8.951e-03  -0.102  0.91887    
after                                     -1.943e-03  8.945e-03  -0.217  0.82802    
income                                    -5.617e-08  7.930e-08  -0.708  0.47871    
age                                       -8.746e-05  7.049e-05  -1.241  0.21466    
political_orientationliberal               8.334e-04  3.069e-03   0.272  0.78594    
political_orientationmoderate             -2.494e-03  2.946e-03  -0.847  0.39727    
treatment:after                            4.350e-02  1.266e-02   3.436  0.00059 ***
treatment:income                          -6.924e-08  1.135e-07  -0.610  0.54183    
treatment:age                              1.317e-04  1.031e-04   1.276  0.20179    
treatment:political_orientationliberal    -4.079e-03  4.295e-03  -0.950  0.34233    
treatment:political_orientationmoderate    2.463e-03  4.130e-03   0.596  0.55092    
after:income                               2.853e-08  1.121e-07   0.254  0.79919    
after:age                                  2.262e-05  9.968e-05   0.227  0.82047    
after:political_orientationliberal         2.752e-04  4.340e-03   0.063  0.94944    
after:political_orientationmoderate        3.176e-03  4.166e-03   0.762  0.44587    
treatment:after:income                     7.008e-07  1.605e-07   4.366 1.27e-05 ***
treatment:after:age                       -6.586e-05  1.459e-04  -0.452  0.65160    
treatment:after:political_orientationliberal  -1.169e-02  6.074e-03  -1.925  0.05427 .  
treatment:after:political_orientationmoderate -1.324e-02  5.841e-03  -2.267  0.02338 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.195 on 103980 degrees of freedom
Multiple R-squared:  0.02858,	Adjusted R-squared:  0.0284 
F-statistic:   161 on 19 and 103980 DF,  p-value: < 2.2e-16
```

Income shows a significant positive interaction, with every $10,000 increase in income leading to a 0.7 percentage point stronger CCPA effect, indicating higher-income customers are more responsive to privacy regulations. Age, contrary to initial hypotheses, shows no significant moderation of the policy's impact ($p = 0.652$), suggesting privacy concerns are consistent across age groups. Political affiliation demonstrates interesting patterns, with both liberal ($p = 0.054$) and moderate ($p = 0.023$) customers showing less positive responses to CCPA compared to conservatives, challenging assumptions about privacy preferences across political ideologies. These findings suggest that income is the primary driver of heterogeneous responses to privacy regulations, while age plays a minimal role and political affiliation shows unexpected patterns.

Based on the analysis, marketing focused actions should focus on two key demographic dimensions. Income-based strategies should prioritize premium privacy features and tiered control options for higher-income segments, as they show stronger response to privacy regulations. While age-based targeting isn't necessary due to consistent effects across age groups, political affiliation requires nuanced messaging: conservative areas respond better to privacy protection emphasis, while liberal/moderate regions need focus on broader product benefits. This suggests a dual approach emphasizing income-based segmentation and politically-aware messaging strategies.

R Code:

```r
# Load required packages
library(tidyverse)

# Load dataset
privacy <- read.csv("privacy.csv")

# Run heterogeneity analysis model
model_hetero <- lm(
  online_ratio ~ treatment * after * (income + age + political_orientation),
  data = privacy
)

# Display coefficient summary
summary(model_hetero)
```

Part III) Interpretation

A. Shown below:

```{r}
data <- read.csv("bank-full.csv", sep=';')
data$y <- ifelse(data$y == "yes", 1, 0)
```

```{r}
glm_model1 <- glm(y ~ duration + campaign, data = data, family = binomial)
summary(glm_model1)
```

```
Call:
glm(formula = y ~ duration + campaign, family = binomial, data = data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.872e+00  3.237e-02  -88.72   <2e-16 ***
duration     3.544e-03  5.534e-05   64.05   <2e-16 ***
campaign    -1.394e-01  9.438e-03  -14.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 32631  on 45210  degrees of freedom
Residual deviance: 27213  on 45208  degrees of freedom
AIC: 27219

Number of Fisher Scoring iterations: 6
```

```{r}
glm_model2 <- glm(y ~ duration + campaign + duration*campaign, data = data, family = binomial)
summary(glm_model2)
```

```
Call:
glm(formula = y ~ duration + campaign + duration * campaign,
    family = binomial, data = data)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.631e+00  4.306e-02 -61.095   <2e-16 ***
duration           3.005e-03  8.177e-05  36.751   <2e-16 ***
campaign          -2.463e-01  1.680e-02 -14.662   <2e-16 ***
duration:campaign  2.235e-04  2.624e-05   8.519   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 32631  on 45210  degrees of freedom
Residual deviance: 27137  on 45207  degrees of freedom
AIC: 27145

Number of Fisher Scoring iterations: 6
```

```{r}
categorical_vars <- c("job", "marital", "education", "default", "housing", "loan", "month")
data[categorical_vars] <- lapply(data[categorical_vars], as.factor)
```

```{r}
glm_model3 <- glm(y ~ duration + campaign + duration*campaign + age + job + marital + education + default +
balance + housing + loan + month,
               data = data, family = binomial)

# Summary of the model
summary(glm_model3)
```

```
Call:
glm(formula = y ~ duration + campaign + duration * campaign + 
    age + job + marital + education + default + balance + housing + 
    loan + month, family = binomial, data = data)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)        -1.976e+00  1.491e-01 -13.255  < 2e-16 ***
duration            3.487e-03  8.728e-05  39.952  < 2e-16 ***
campaign           -2.094e-01  1.715e-02 -12.212  < 2e-16 ***
age                 1.675e-03  2.072e-03   0.808 0.418862    
jobblue-collar     -4.023e-01  6.921e-02  -5.812 6.16e-09 ***
jobentrepreneur    -5.170e-01  1.210e-01  -4.271 1.94e-05 ***
jobhousemaid       -5.966e-01  1.306e-01  -4.567 4.95e-06 ***
jobmanagement      -2.100e-01  6.953e-02  -3.021 0.002523 ** 
jobretired          2.309e-01  9.100e-02   2.537 0.011188 *  
jobself-employed   -3.603e-01  1.066e-01  -3.379 0.000728 ***
jobservices        -2.843e-01  7.983e-02  -3.561 0.000370 ***
jobstudent          5.046e-01  1.028e-01   4.907 9.27e-07 ***
jobtechnician      -2.100e-01  6.539e-02  -3.211 0.001322 ** 
jobunemployed      -1.961e-01  1.047e-01  -1.872 0.061143 .  
jobunknown         -4.653e-01  2.191e-01  -2.124 0.033683 *  
maritalmarried     -1.487e-01  5.582e-02  -2.663 0.007734 ** 
maritalsingle       1.519e-01  6.369e-02   2.386 0.017046 *  
educationsecondary  2.351e-01  6.176e-02   3.807 0.000140 ***
educationtertiary   5.190e-01  7.150e-02   7.259 3.90e-13 ***
educationunknown    2.420e-01  9.860e-02   2.454 0.014121 *  
defaultyes         -1.941e-01  1.590e-01  -1.221 0.221964    
balance             1.479e-05  4.765e-06   3.104 0.001912 ** 
housingyes         -7.733e-01  4.153e-02 -18.617  < 2e-16 ***
loanyes            -5.276e-01  5.806e-02  -9.087  < 2e-16 ***
monthaug           -7.746e-01  7.412e-02 -10.450  < 2e-16 ***
monthdec            9.315e-01  1.626e-01   5.730 1.00e-08 ***
monthfeb           -2.417e-01  8.047e-02  -3.004 0.002663 ** 
monthjan           -1.145e+00  1.147e-01  -9.988  < 2e-16 ***
monthjul           -9.335e-01  7.351e-02 -12.699  < 2e-16 ***
monthjun           -7.821e-01  7.504e-02 -10.422  < 2e-16 ***
monthmar            1.486e+00  1.137e-01  13.069  < 2e-16 ***
monthmay           -1.067e+00  6.571e-02 -16.235  < 2e-16 ***
monthnov           -8.785e-01  8.059e-02 -10.900  < 2e-16 ***
monthoct            9.273e-01  9.905e-02   9.363  < 2e-16 ***
monthsep            1.006e+00  1.081e-01   9.307  < 2e-16 ***
duration:campaign   2.019e-04  2.642e-05   7.643 2.13e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 32631  on 45210  degrees of freedom
Residual deviance: 23723  on 45175  degrees of freedom
AIC: 23795

Number of Fisher Scoring iterations: 6
```

B. Duration (**duration** = 3.487e-03, p < 2e-16) - This coefficient represents the log-odds increase in y (subscribing to the term deposit) for every additional unit of call duration. Exponentiating the coefficient: $e^{0.003487} \approx 1.00349$. This means that for every one-unit increase in call duration, the odds of a successful campaign (y=1) increase by about 0.35%. Since the p-value is extremely low, this effect is statistically significant.

Campaign (**campaign** = -2.094e-01, p < 2e-16) - This coefficient represents the log-odds change in y for each additional contact attempt. Exponentiating the coefficient: $e^{-0.2094} \approx 0.811$. This suggests that each additional contact attempt decreases the odds of success by about 18.9%. The negative coefficient means that increasing the number of contacts reduces the likelihood of a positive response, potentially due to customer annoyance or diminishing returns.

Interaction Term (**duration:campaign** = 2.019e-04, p < 2.13e-14) - This term captures how the effect of duration changes depending on the number of contact attempts. Since the coefficient is positive, it indicates that the negative impact of increasing contacts (campaign) is mitigated when call duration is longer. Exponentiating the coefficient: $e^{0.0002019} \approx 1.0002$. The increase is small, but statistically significant. This suggests that longer calls reduce the negative effect of frequent contacts, meaning that if the conversation lasts longer, multiple attempts may not be as harmful.

Overall, we can say that call duration is a strong predictor of success, meaning longer calls significantly increase the likelihood of a positive response. Too many contacts reduce success, i.e. aggressively reaching out to customers multiple times can backfire. However, longer calls compensate for the downsides of frequent contact - if an agent manages to have a meaningful conversation (longer duration), the negative effect of multiple attempts is minimized.

These insights can help the bank refine its marketing strategy:

- Focus on longer, more meaningful conversations rather than repeatedly calling customers.
- Avoid excessive contact attempts, as diminishing returns set in quickly.
- If multiple contacts are necessary, ensure that conversations are engaging and valuable to the customer.

C. The coefficients represent the log-odds of the response variable (accepting the campaign offer) relative to the reference category (admin). Positive coefficients (students, retired folks) indicate a higher likelihood of accepting the campaign offer compared to the reference category. Negative coefficients indicate a lower likelihood of accepting the campaign offer compared to the reference category.

Ranking of Job Titles (in terms of coefficients, Statistical significance will be determined using t statistic value)

Based on the coefficients, we can rank the job titles in terms of their likelihood of accepting the campaign offer (from highest to lowest likelihood):

1. **jobstudent** (Coefficient: 0.5046) - Students are the most likely to accept the campaign offer.
2. **jobretired** (Coefficient: 0.2309) - Retired individuals are also more likely to accept the offer compared to the reference category.
3. **Unemployed** (-0.1961) - This group is next in the ranking list, compared to the reference category.
4. **jobmanagement** (Coefficient: -0.2100) - Managers are less likely to accept the offer compared to students and retired individuals but are still relatively higher than other job titles.
5. **jobtechnician** (Coefficient: -0.2100) - Technicians have the same coefficient as managers, indicating a similar likelihood of accepting the offer.
6. **jobservices** (Coefficient: -0.2843) - Individuals in services are less likely to accept the offer compared to managers and technicians.
7. **jobself-employed** (Coefficient: -0.3603) - Self-employed individuals are even less likely to accept the offer.
8. **jobblue-collar** (Coefficient: -0.4023) - Blue-collar workers are less likely to accept the offer compared to self-employed individuals.
9. **jobunknown** (Coefficient: -0.4653) - Individuals with unknown job titles are less likely to accept the offer.
10. **jobentrepreneur** (Coefficient: -0.5170) - Entrepreneurs are less likely to accept the offer compared to most other job titles.
11. **jobhousemaid** (Coefficient: -0.5966) - Housemaids are the least likely to accept the campaign offer.

Now, we use the t-statistic formula (difference between coefficients divided by sqrt of sum of squared standard errors) and check which all differences in job title likelihood are statistically significant at the 95% confidence level ($p < 0.05$). Wherever t statistic > 1.96, the difference is significant. After doing these calculations manually, we arrive at the following major conclusions-

- Students are significantly more likely to accept the campaign compared to all other job categories.
- Retired individuals also show a significantly higher likelihood than most working professionals.
- Housemaids and Entrepreneurs are significantly less likely to accept the campaign compared to other job roles.
- The difference between Self-Employed vs. Technicians, Blue-Collar vs. Services is not statistically significant, suggesting similar response rates.

D. We cannot claim causality in the relationship between duration or campaign and the response because this analysis is based on observational data rather than a controlled experiment. The relationships identified in the regression analysis indicate correlation, not causation. There may be confounding variables that influence both call duration, the number of contacts, and the customer's response, leading to biased estimates in our model.

One significant confound that is not included in the dataset is customer interest level in the bank's offer. Customers who are already interested in the bank's services may be more willing to stay on a call longer or be more receptive to multiple contacts. Conversely, those who are not interested may end the call quickly or ignore repeated calls altogether. This confound directly correlates with both call duration and the number of contact attempts, making it an essential missing factor.

The customer's interest level also affects their likelihood of accepting the campaign offer. A highly interested customer is more likely to accept the term deposit, whereas an uninterested customer is more likely to reject it, regardless of how long the call lasts or how many times they are contacted. This creates a suspicious correlation where longer call duration or frequent contact appears to drive higher acceptance rates, when in reality, pre-existing interest is the underlying cause.

To obtain better estimates of causality, the bank should attempt to collect data on customer interest levels. This could be done by tracking past interactions, such as whether a customer previously inquired about investment products, their browsing behavior on the banking app, or their engagement with marketing emails. Additionally, survey data about customer preferences could provide insights into who is genuinely interested in term deposits. By adding such data, the model could more accurately differentiate between genuine marketing effects and pre-existing customer interest.

To rigorously assess causality, the bank should consider A/B testing, where some customers are randomly assigned different call durations or contact frequencies. This would help isolate true causal effects by ensuring that duration and campaign frequency are not confounded by customer interest. Another approach could be using instrumental variables, i.e. factors that influence duration or campaign frequency but are not directly tied to customer interest - to better estimate the actual impact.

E. To establish causality between duration and campaign frequency on customer response (y), the bank should conduct Randomized Controlled Trials (RCTs). By randomly assigning customers to different conditions, we can eliminate confounding factors and ensure that changes in response rate are due to variations in call duration or campaign frequency rather than external factors.

RCT for Call Duration

To test the causal effect of call duration on the likelihood of a positive response, the bank should randomly assign customers to different call length groups. The experiment should proceed as follows:

1. Random Assignment: Customers who are contacted receive calls of randomly assigned durations, such as 30 seconds, 1 minute, 2 minutes, and 3+ minutes.
2. Standardized Call Script: To ensure that only duration varies, all agents must follow a fixed script without introducing additional persuasive efforts.
3. Control for Customer Interest: The randomization ensures that customers across different call duration groups have the same average level of prior interest.
4. Measure Response Rates: The percentage of customers in each group who accept the offer will indicate whether longer calls increase acceptance.
5. Statistical Analysis: A difference in means test (t-test or ANOVA) can determine whether longer calls significantly improve response rates.

If we find that longer calls increase conversion rates, we can conclude that call duration probably has a causal impact on campaign success. If not, it suggests that merely increasing call length does not drive customer decisions.

RCT for Campaign Frequency

To assess the causal impact of number of contact attempts (campaign frequency) on response rates, the bank should randomly assign customers to different contact frequencies:

1. Randomized Contact Frequency: Customers are assigned to different groups, such as 1 call, 3 calls, and 5 calls within a given time period.

2. Time Gap Between Calls: Ensure that the time between contact attempts is standardized (e.g., 3-day intervals) to avoid bias from customers who may naturally become more receptive over time.
3. Consistent Messaging: To isolate the impact of frequency, all calls should deliver the same script and be made by trained agents.
4. Measure Drop-off Rate: If a customer rejects the offer after one call, they should still be included in follow-ups based on their assigned group.
5. Statistical Evaluation: Compare conversion rates across groups to determine if more contact attempts increase or decrease acceptance rates.

If the 3-contact group converts better than the 1-contact group but the 5-contact group converts worse, it may suggest that persistence helps up to a point but excessive contact may backfire.

The bank can implement these RCTs by integrating them into its call center system, ensuring that randomization occurs automatically when a customer is selected for a campaign. One challenge is ethical considerations, as some customers may find excessive contact annoying. Thus, opt-out options should be available.

By running these RCTs, the bank can attempt to establish a causal relationship between duration and campaign frequency on customer response, allowing for data-driven optimization of future marketing efforts.

F.

````
```{r}
alternative_1 <- data
alternative_1$campaign <- ifelse(alternative_1$poutcome == "failure", round(alternative_1$campaign * 0.7),
alternative_1$campaign)

alternative_2 <- data
alternative_2$duration <- ifelse(alternative_2$poutcome == "success", round(alternative_2$duration * 1.2),
alternative_2$duration)
```
````

````
```{r}
glm_model_alt1 <- glm(y ~ duration + campaign + duration*campaign + age + job + marital + education + default
+ balance + housing + loan + month,
                  data = alternative_1, family = binomial)

glm_model_alt2 <- glm(y ~ duration + campaign + duration*campaign + age + job + marital + education + default
+ balance + housing + loan + month,
                  data = alternative_2, family = binomial)
```
````

```{r}
predicted_y_alt1 <- predict(glm_model_alt1, newdata = alternative_1, type = "response")
predicted_y_alt2 <- predict(glm_model_alt2, newdata = alternative_2, type = "response")

avg_y_original <- mean(predict(glm_model3, newdata = data, type = "response"))
avg_y_alt1 <- mean(predicted_y_alt1)
avg_y_alt2 <- mean(predicted_y_alt2)

diff_alt1 <- avg_y_alt1 - avg_y_original
diff_alt2 <- avg_y_alt2 - avg_y_original

if (diff_alt1 > diff_alt2) {
  best_alternative <- "Reducing campaign calls for past failures is more effective"
} else {
  best_alternative <- "Increasing call duration for past successes is more effective"
}

print(paste("Best alternative:", best_alternative))
```

```
[1] "Best alternative: Reducing campaign calls for past failures is more effective"
```

Data	
alternative_1	45211 obs. of 17 variables
alternative_2	45211 obs. of 17 variables
data	45211 obs. of 17 variables
glm_model_alt1	List of 30
glm_model_alt2	List of 30
glm_model1	List of 30
glm_model2	List of 30
glm_model3	List of 30
Values	
avg_y_alt1	0.116984806493255
avg_y_alt2	0.116984806435646
avg_y_original	0.116984806179195
best_alternative	"Reducing campaign calls for past failures i…
categorical_vars	chr [1:7] "job" "marital" "education" "defau…
diff_alt1	3.14060291661811e-10
diff_alt2	2.56451582192341e-10
predicted_y_alt1	Large numeric (45211 elements, 3.3 MB)
predicted_y_alt2	Large numeric (45211 elements, 3.3 MB)