Empirical Economics Final Project

Does Previous Contact Influence Subscription Rates?

Team 32
Danielle Dawazhuoma
Shrishti Agarwal
Utkarsh Gupta
Lizzie Wang
Zheng Xie

Motivation

Banking & Marketing Relevance

- Many banks rely on **direct marketing campaigns** to acquire new customers.
- Understanding whether previous contact influences subscription likelihood **allows banks to optimize marketing strategies** and **reduce unnecessary customer outreach costs**.

Economic & Business Importance

- **Direct marketing costs are high**, and inefficient customer targeting can be costly.
- If previous contact has a strong impact on customer behavior, firms can improve ROI (Return on Investment) by **focusing on the most responsive customers**.

Customer Behavior & Decision-Making

- **Not all customers respond positively to repeated outreach**; some may become disengaged or even annoyed.
- If previous contact increases conversion rates, banks should **prioritize follow-up calls**.
- If previous contact has no effect or even discourages some customers, banks should **change their marketing strategy.**

Literature

What previous research tells us

Existing Studies in Financial Marketing

- Moro et al. (2011) used **logistic regression models** to analyze the impact of direct marketing on term deposit subscriptions.
- Their findings suggest that **direct engagement increases customer conversion rates**, but they did not account for **selection bias**.
- Other studies (Ding et al., 2019) warn that **too many contacts may reduce engagement**.

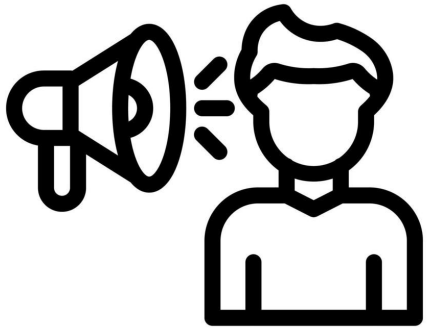
Limitations of Prior Work

- **Most studies rely on simple regression models (OLS, Logistic Regression)** which **do not fully address causality**.
- They often fail to control for **self-selection bias**, where high-intent customers may **already be more likely to subscribe**, making the effect of contact difficult to isolate.

Literature

Our Contribution

- Applied **multiple causal inference methods** (PSM, Regression Adjustment, DiD) to **eliminate selection bias and improve causal validity**.
- Used **machine learning methods** (Meta Learners, Causal Random Forests) to **estimate heterogeneous treatment effects** (HTEs).
- Unlike prior studies, we identify **which customer segments benefit most from previous contact**, allowing for more targeted marketing strategies.
- Combined econometric and machine learning approaches to provide a **comprehensive, data-driven framework** for customer engagement optimization.



The Dataset

Dataset Overview

- The dataset captures direct marketing campaigns conducted by a **Portuguese banking institution,** where customers were **contacted via phone calls to promote term deposit subscriptions**.which contains 4,521 observations on customer demographics, past interactions, and marketing campaign outcomes.
- The dataset originates from Moro et al. (2011), a study analyzing the effectiveness of direct marketing campaigns in the banking sector.
- The key objective is to predict whether a customer subscribes to a term deposit after being contacted.

The Dataset - Key Variables

Treatment Variable: Previous contacts (previous > 0 vs. previous = 0).

Outcome Variable: Subscription status (y = yes/no).

Control Variables:

- **Customer attributes**: Age, job type, marital status, education, financial status (balance, loans).
- **Campaign interactions**: Number of contacts (campaign), contact type (telephone/cellular), call duration (duration).
- **Past interactions**: Days since last contact (pdays), success of past campaigns (poutcome).

Why This Data?

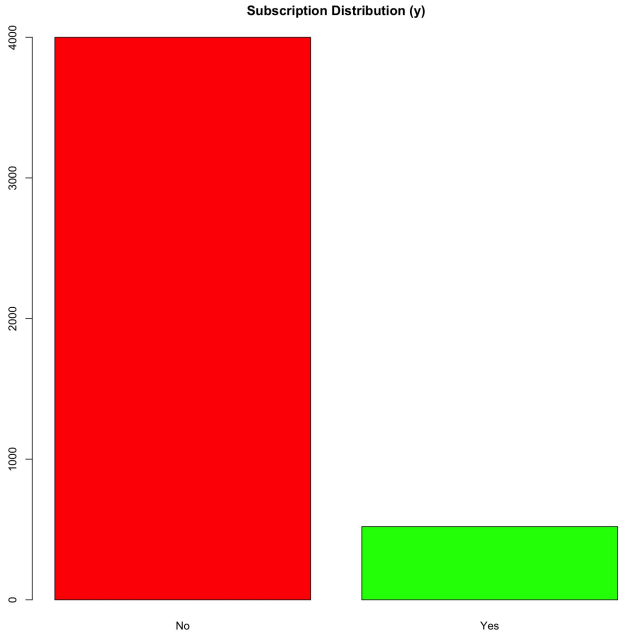
- Provides a **rich set of features** to control for potential confounders.
- Allows us to **apply multiple causal inference techniques** to estimate the true impact of past contacts on subscription rates.
- Enables the use of **both econometric and machine learning approaches** to uncover heterogeneous treatment effects.

Exploratory Data Analysis (EDA)

```
> print(summary(df))
      age                job         marital       education       default        balance       housing      loan
 Min.   :19.00   management :969   divorced: 528   primary  : 678   Min.   :0.00000   Min.   :-3313   Min.   :0.000   Min.   :0.0000
 1st Qu.:33.00   blue-collar:946   married :2797   secondary:2306   1st Qu.:0.00000   1st Qu.:   69   1st Qu.:0.000   1st Qu.:0.0000
 Median :39.00   technician :768   single  :1196   tertiary :1350   Median :0.00000   Median :  444   Median :1.000   Median :0.0000
 Mean   :41.17   admin.     :478                   unknown  : 187   Mean   :0.01681   Mean   : 1423   Mean   :0.566   Mean   :0.1528
 3rd Qu.:49.00   services   :417                                     3rd Qu.:0.00000   3rd Qu.: 1480   3rd Qu.:1.000   3rd Qu.:0.0000
 Max.   :87.00   retired    :230                                     Max.   :1.00000   Max.   :71188   Max.   :1.000   Max.   :1.0000
                 (Other)    :713
     contact           day             month          duration       campaign          pdays          previous        poutcome         y
 cellular :2896   Min.   : 1.00   may    :1398   Min.   :   4   Min.   : 1.000   Min.   : -1.00   Min.   : 0.0000   failure: 490   Min.   :0.0000
 telephone: 301   1st Qu.: 9.00   jul    : 706   1st Qu.: 104   1st Qu.: 1.000   1st Qu.: -1.00   1st Qu.: 0.0000   other  : 197   1st Qu.:0.0000
 unknown  :1324   Median :16.00   aug    : 633   Median : 185   Median : 2.000   Median : -1.00   Median : 0.0000   success: 129   Median :0.0000
                  Mean   :15.92   jun    : 531   Mean   : 264   Mean   : 2.794   Mean   : 39.77   Mean   : 0.5426   unknown:3705   Mean   :0.1152
                  3rd Qu.:21.00   nov    : 389   3rd Qu.: 329   3rd Qu.: 3.000   3rd Qu.: -1.00   3rd Qu.: 0.0000                  3rd Qu.:0.0000
                  Max.   :31.00   apr    : 293   Max.   :3025   Max.   :50.000   Max.   :871.00   Max.   :25.0000                  Max.   :1.0000
                                  (Other): 571
```

- **Subscription Rate**: The subscription rate is relatively low, with only 11.5% of customers subscribing. It seems that factors such as prior contact and customer characteristics (e.g., loan, housing) could play a role in influencing this outcome.
- **Previous Contact**: The variable "previous" indicates that a large portion of the customers were contacted multiple times.
- **Contact Method and Outcome**: Customers contacted by cellular phone are more common, and the outcome of previous campaigns appears mixed, with a fairly high number of failures.

Exploratory Data Analysis (EDA)



The bar plot illustrates the distribution of subscription outcomes in the dataset. The majority of customers did not subscribe (shown in red), while a smaller proportion did subscribe (shown in green).

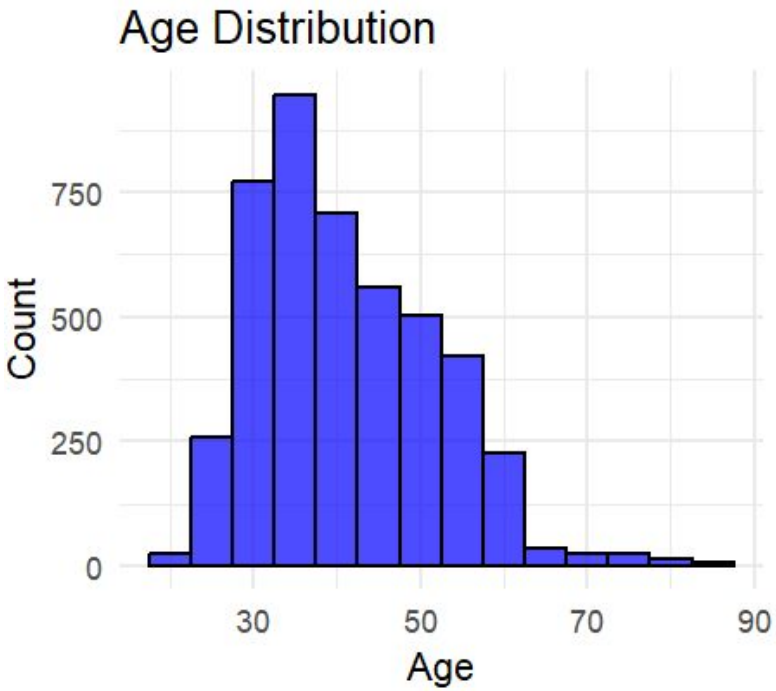
Subscription Rate Calculation:

- Total Customers: 4,521
- Subscribed (Yes): 521 (~**11.53%**)
- Not Subscribed (No): 4,000 (~88.47%)

Interpretation:

- Highly Imbalanced Data → A large percentage of customers declined the offer.
- Marketing Implications → Since only 11.53% of customers converted, understanding which factors influence subscriptions is critical.

Exploratory Data Analysis (EDA)



The histogram displays the age distribution of customers in the dataset. The majority of customers fall between the 30-50 age range, with the highest concentration in the 30-35 bracket.

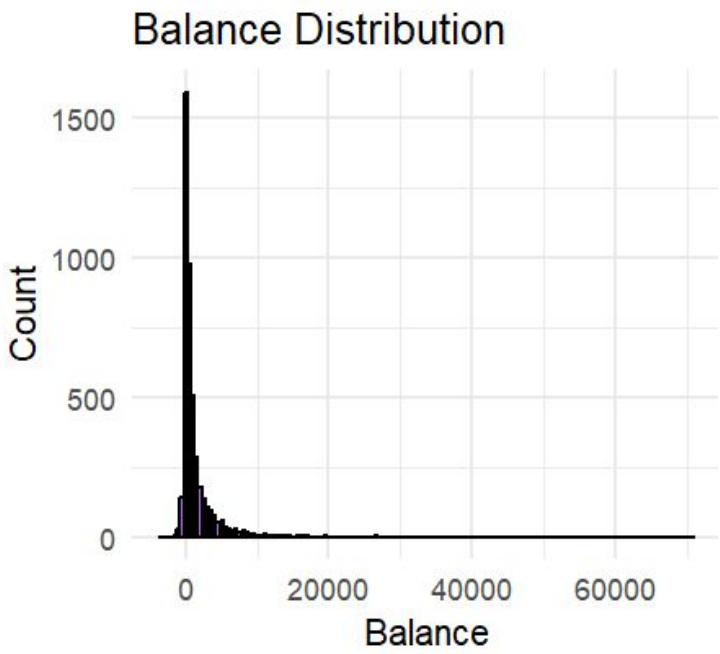
Age Breakdown:

- **Young Adults** (15-25): **1.5%** of customers
- **Early Career** (25-35): **31%** of customers (largest segment)
- **Mid Career** (35-50): **43%** of customers
- **Older Adults** (50-70): **21%** of customers
- **Seniors** (70+): **1.3%** of customers

Interpretation:

- The distribution is right-skewed, meaning **most customers are in their 30s-50s**, with fewer older customers.
- This indicates that the bank's marketing campaigns primarily target **working-age individuals**.

Exploratory Data Analysis (EDA)



The histogram shows a **highly skewed distribution of customer average yearly balances (euros)**. A significant proportion of customers have very low balances, while a small number of customers have extremely high balances (outliers).

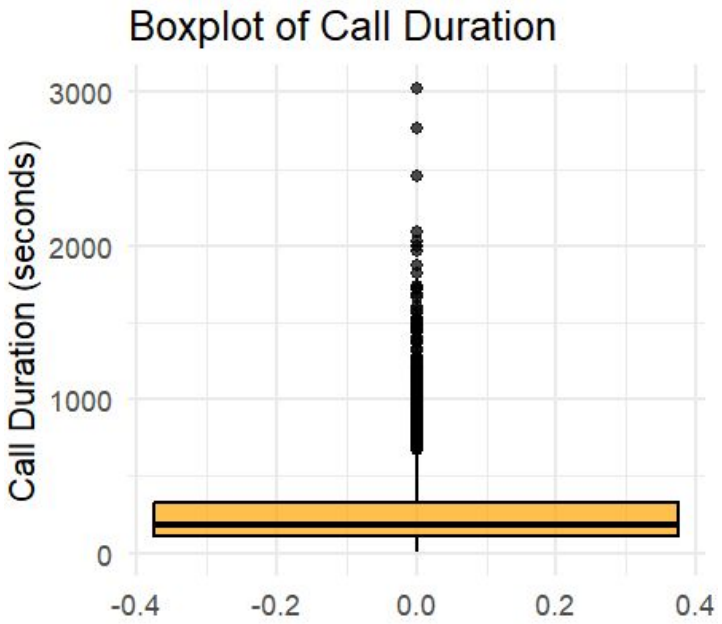
Balance Breakdown:

- **Majority of customers (~55%) have balances between -313 and 687**.
- A small group (~25%) has balances above 1,690, indicating a wealthier segment.
- Outliers exist, with some customers having balances exceeding 60,000, but they are rare.

Interpretation:

- Subscription likelihood may vary by balance, meaning we should analyze whether higher balances correlate with higher term deposit subscriptions.

Exploratory Data Analysis (EDA)



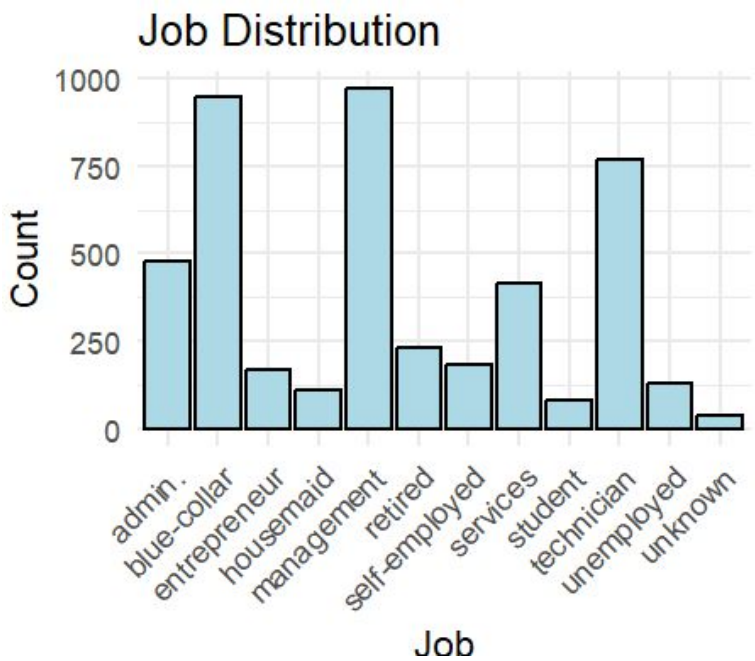
The boxplot shows that most call durations are concentrated within a short range, but there are many extreme outliers.

- **Median call duration is 185 seconds (~3 minutes)**, meaning half of the calls lasted less than this.
- The interquartile range (IQR) is 225 seconds, indicating that **most calls lasted between 104 and 329 seconds** (~1.7 to 5.5 minutes).
- The **maximum call duration is 3025 seconds (~50 minutes)**, highlighting a small number of exceptionally long calls.

Interpretation:

- The right-skewed distribution suggests that while **most calls are short, some last significantly longer**.
- Longer calls could indicate more engaged conversations, potentially leading to higher subscription rates.

Exploratory Data Analysis (EDA)



The three most common occupations are **management** (969), **blue-collar** (946), and **technician** (768).

- Other large segments include **admin** (478) and **services** (417).
- **Smaller job categories** include students (84), housemaids (112), and unknown (38).

Interpretation:

- The dataset consists primarily of w**orking professionals, with a mix of skilled and unskilled labor.**
- Retired individuals (230) and self-employed workers (183) make up a notable segment, potentially with different financial behaviors.
- Understanding job types can help determine subscription probability, as **income stability and job type may influence financial decisions.**

Exploratory Data Analysis (EDA)

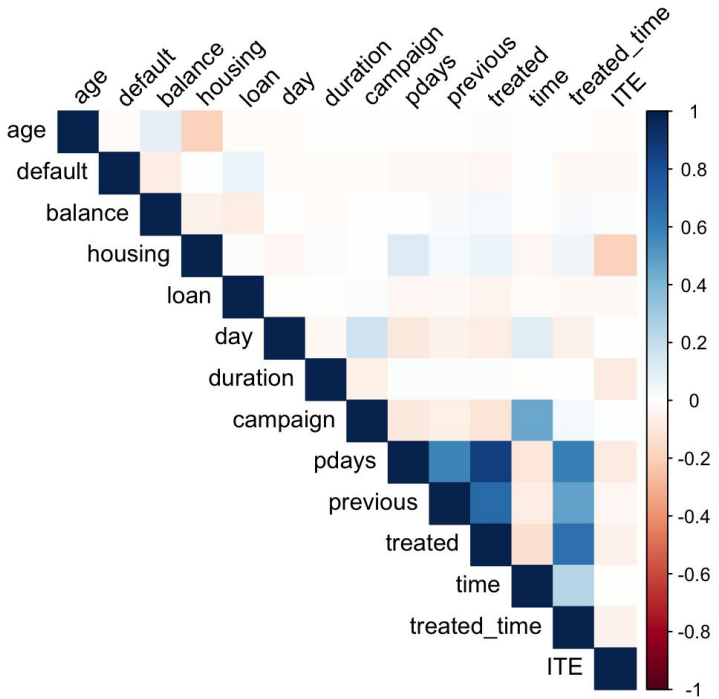
Correlation Analysis:

Call duration is positively correlated with subscription likelihood (ITE), reinforcing the idea that longer conversations improve conversion rates.

Previous contacts (pdays & previous) are moderately correlated with ITE, suggesting past interactions impact futur decisions.

Treatment variables (treated, time, treated_time) show expect correlations, validating our experimental setup.

Balance has weak correlations with most features, indicating that account balance alone may not drive subscription decisions.



METHODOLOGY

Challenges in Estimating Causal Effects

Potential Omitted Variable Bias

- Customers who were previously contacted may have been targeted based on their likelihood of subscribing, not randomly selected.
- Certain factors, like customer income, risk tolerance, or financial literacy, may affect both contact probability and subscription likelihood.

Selection Bias

- Simply comparing subscription rates between contacted and non-contacted customers could be misleading because these two groups may be systematically different

METHODOLOGY

How We Addressed These Issues?

- **OLS Regression (Baseline) – Simple Correlation Analysis**

  OLS regression provides an initial estimate of the relationship between previous contacts and subscription likelihood. However, it only measures correlation and does not account for selection bias or confounding variables, limiting its causal validity.

- **Propensity Score Matching (PSM) – Creating a Balanced Comparison**

  PSM helps create a more comparable control group by matching treated and untreated customers with similar characteristics. This reduces selection bias and makes the treatment effect estimate more reliable, though it does not account for unobserved confounders.

- **Regression Adjustment – Controlling for Confounding Variables**

  By incorporating control variables such as age, balance, and call duration, regression adjustment helps isolate the true impact of previous contacts on subscriptions. While more accurate than OLS alone, it still depends on the assumption that all key confounders are included.

METHODOLOGY

How We Addressed These Issues?

- **Difference-in-Differences (DiD) – Analyzing Changes Over Time**

  DiD compares changes in subscription behavior before and after treatment, using a control group to account for external factors. This method is useful for policy changes or time-based interventions but was less applicable here due to a lack of clear time-based variation.

- **Meta Learners (T-Learner) – Estimating Individual Treatment Effects**

  Meta learners estimate how treatment effects vary across different customers by training separate models for treated and control groups. This allows us to see which types of customers benefit the most from past contacts, providing valuable personalized insights.

- **Causal Random Forests (CRF) – Identifying Key Drivers of Treatment Effects**

  CRF captures nonlinear relationships and ranks the most important features influencing subscription decisions. This method is particularly useful for understanding customer heterogeneity and uncovering complex interactions between variables.

METHODOLOGY - Baseline OLS Model

$$Subscription = \alpha + \beta(PreviousContact) + Controls + \varepsilon$$

```
### OLS Model: Effect of Previous Marketing Contact on Subscription
# Treatment: Clients who were contacted before (`previous > 0`)
# Control: Clients who were never contacted before (`previous = 0`)
model_previous <- lm(y ~ previous + age + balance + duration + job + marital +
                       education + housing + loan, data = df)
summary(model_previous)
```

Baseline OLS Model

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.996e-02  3.421e-02   0.584  0.55950
previous            2.010e-02  2.520e-03   7.973 1.94e-15 ***
age                 7.194e-04  5.229e-04   1.376  0.16891
balance             3.459e-07  1.431e-06   0.242  0.80907
duration            4.938e-04  1.643e-05  30.061  < 2e-16 ***
jobblue-collar     -5.303e-02  1.686e-02  -3.145  0.00167 **
jobentrepreneur    -5.901e-02  2.626e-02  -2.247  0.02466 *
jobhousemaid       -3.991e-02  3.100e-02  -1.287  0.19811
jobmanagement      -2.903e-02  1.834e-02  -1.583  0.11355
jobretired          5.303e-02  2.552e-02   2.078  0.03775 *
jobself-employed   -4.483e-02  2.540e-02  -1.765  0.07762 .
jobservices        -3.455e-02  1.919e-02  -1.800  0.07195 .
jobstudent          6.152e-02  3.510e-02   1.753  0.07972 .
jobtechnician      -3.248e-02  1.680e-02  -1.933  0.05327 .
jobunemployed      -6.850e-02  2.870e-02  -2.387  0.01704 *
jobunknown          3.965e-02  4.935e-02   0.803  0.42176
maritalmarried     -3.022e-02  1.369e-02  -2.207  0.02734 *
maritalsingle      -8.572e-03  1.600e-02  -0.536  0.59216
educationsecondary  6.303e-03  1.390e-02   0.454  0.65015
educationtertiary   3.720e-02  1.697e-02   2.192  0.02840 *
educationunknown   -2.171e-02  2.443e-02  -0.888  0.37443
housing            -5.606e-02  9.122e-03  -6.145 8.68e-10 ***
loan               -5.261e-02  1.195e-02  -4.404 1.08e-05 ***
```

Findings:

- Customers who were **previously contacted are ~11% more likely to subscribe** ($p < 0.001$).
- However, the low R^2 (0.219) suggests that **other factors also play a role in determining subscription likelihood.**

Limitations:

- Selection bias remains unaddressed—OLS does not account for why certain customers were contacted.
- Unobserved factors (e.g., risk preference) may still bias estimates.

METHODOLOGY - Propensity Score Matching

Why use PSM?

- PSM is a statistical method used to reduce selection bias in observational studies. It creates a **balanced comparison** between the treated and untreated groups by matching customers with similar probabilities (propensities) of receiving the treatment, based on their observed characteristics.

- PSM ensures that the comparison between treated and untreated customers is based on similar underlying characteristics. This reduces the influence of confounding variables and provides a more reliable estimate of the treatment effect.

- By creating a balanced comparison, PSM allows us to isolate the effect of previous contact on subscription rates, making the results more interpretable as causal.

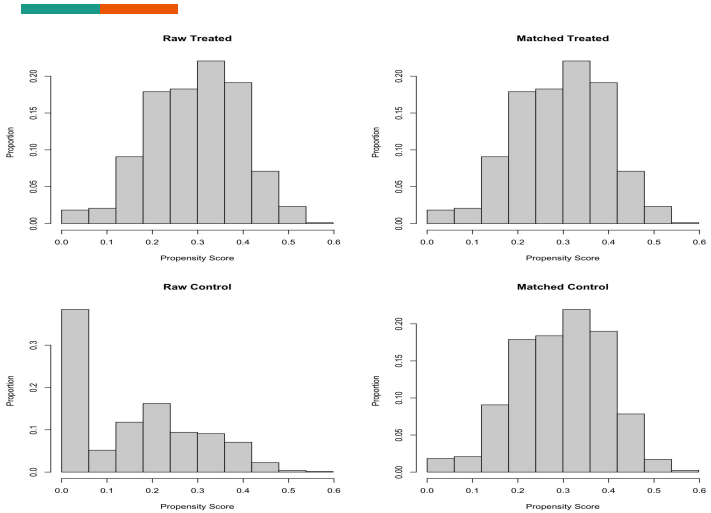
Propensity Score Matching - R Code

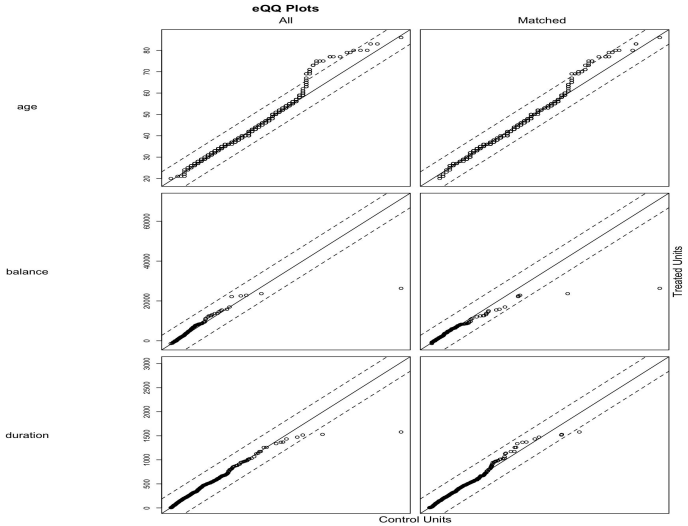
```r
# Load necessary libraries
library(dplyr)
library(MatchIt)
library(ggplot2)

# Define treatment and control groups
df <- df %>%
  mutate(treated = ifelse(previous > 0, 1, 0))  # 1 = contacted before, 0 = never contacted

# Define covariates for matching (excluding 'previous' to prevent bias)
covariates <- c("age", "balance", "duration", "campaign", "housing", "loan", "job", "marital", "education", "contact")

# Estimate propensity score using logistic regression
ps_model <- glm(treated ~ age + balance + duration + campaign + housing + loan + job + marital + education + contact,
                family = binomial(), data = df)

# Perform Propensity Score Matching (Nearest Neighbor, 1:1 Matching)
matched_data <- matchit(treated ~ age + balance + duration + campaign + housing + loan + job + marital + education + contact,
                        data = df, method = "nearest", ratio = 1)

# Summary of matching
summary(matched_data)

# Create a dataframe with matched observations
df_matched <- match.data(matched_data)

# Check balance before and after matching
plot(matched_data, type = "hist")  # Histogram of propensity scores
plot(matched_data, type = "qq")    # Q-Q plot of covariate balance

# Estimate treatment effect (ATE & ATT) using OLS on matched data
model_psm <- lm(y ~ treated, data = df_matched)
summary(model_psm)

# Convert treated variable to a factor for categorical plotting
df_matched$treated <- as.factor(df_matched$treated)

# Calculate mean subscription rate for each group
subscription_summary <- df_matched %>%
  group_by(treated) %>%
  summarise(mean_subscription = mean(y))
```

```
Summary of Balance for Matched Data:
                Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max Std. Pair Dist.
distance               0.2936        0.2932          0.0044     1.0140    0.0004    0.0147          0.0054
age                   41.5625       41.5135          0.0043     1.1283    0.0101    0.0404          1.0633
balance             1639.6716     1676.9510         -0.0124     0.5075    0.0258    0.0821          0.8244
duration             272.3983      267.7745          0.0188     1.0114    0.0098    0.0404          0.9472
campaign               2.0159        1.9988          0.0108     1.3511    0.0051    0.0613          0.6307
housing                0.6324        0.6483         -0.0330          .    0.0159    0.0159          0.5210
loan                   0.1213        0.1348         -0.0413          .    0.0135    0.0135          0.6719
jobadmin.              0.1324        0.1324          0.0000          .    0.0000    0.0000          0.2010
jobblue-collar         0.1863        0.1789          0.0189          .    0.0074    0.0074          0.7617
jobentrepreneur        0.0282        0.0331         -0.0296          .    0.0049    0.0049          0.3702
jobhousemaid           0.0221        0.0270         -0.0334          .    0.0049    0.0049          0.3004
jobmanagement          0.2279        0.2230          0.0117          .    0.0049    0.0049          0.8296
jobretired             0.0600        0.0576          0.0103          .    0.0025    0.0025          0.4642
jobself-employed       0.0343        0.0380         -0.0202          .    0.0037    0.0037          0.3837
jobservices            0.0760        0.0846         -0.0324          .    0.0086    0.0086          0.5041
jobstudent             0.0270        0.0257          0.0076          .    0.0012    0.0012          0.3253
jobtechnician          0.1728        0.1679          0.0130          .    0.0049    0.0049          0.7390
jobunemployed          0.0257        0.0245          0.0077          .    0.0012    0.0012          0.2864
jobunknown             0.0074        0.0074          0.0000          .    0.0000    0.0000          0.0147
maritaldivorced        0.1042        0.1164         -0.0401          .    0.0123    0.0123          0.6258
maritalmarried         0.6029        0.6127         -0.0200          .    0.0098    0.0098          0.9918
maritalsingle          0.2929        0.2708          0.0485          .    0.0221    0.0221          0.9102
educationprimary       0.1201        0.1066          0.0415          .    0.0135    0.0135          0.6296
educationsecondary     0.5086        0.5270         -0.0368          .    0.0184    0.0184          0.9340
educationtertiary      0.3260        0.3321         -0.0131          .    0.0061    0.0061          0.9124
educationunknown       0.0453        0.0343          0.0530          .    0.0110    0.0110          0.3475
contactcellular        0.8971        0.8934          0.0121          .    0.0037    0.0037          0.5364
contacttelephone       0.0895        0.0919         -0.0086          .    0.0025    0.0025          0.5668
contactunknown         0.0135        0.0147         -0.0106          .    0.0012    0.0012          0.0106

Sample Sizes:
          Control Treated
All          3705     816
Matched       816     816
Unmatched    2889       0
Discarded       0       0
```

The dataset is well-balanced, with 816 treated (previously contacted) and 816 control (never contacted) units. This ensures that the treatment and control groups are comparable in terms of observed covariates

Propensity Score Matching





Findings:

The **balance diagnostics** (histograms and Q-Q plots) confirm that the distributions of covariates are similar between the two groups after matching, reducing the risk of selection bias.

Propensity Score Matching

```
Call:
lm(formula = y ~ treated, data = df_matched)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2255 -0.2255 -0.1140 -0.1140  0.8860

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.11397    0.01300   8.764  < 2e-16 ***
treated      0.11152    0.01839   6.064 1.65e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3715 on 1630 degrees of freedom
Multiple R-squared:  0.02206,   Adjusted R-squared:  0.02146
F-statistic: 36.77 on 1 and 1630 DF,  p-value: 1.645e-09
```

Findings :

- The **estimated treatment effect** (coefficient of treated in the linear regression model) is **0.11152**. This means that customers who were previously contacted are, on average, **11.15 percentage points more likely to subscribe** to a term deposit compared to those who were never contacted.
- The effect is **highly statistically significant** (p-value = **1.65e-09**), indicating strong evidence that previous marketing contact positively impacts subscription behavior.
- **The Adjusted R-squared value (0.02146) is low**, meaning that while previous contact has a significant effect, other factors likely influence the subscription decision as well.

Propensity Score Matching

Result:

The analysis support the **causal claim** that **previous contact increases the likelihood of subscription**. This implies that banks can improve their marketing ROI by focusing on customers who have been contacted before, as they are more likely to respond positively.

Limitations :

- PSM only accounts for observed confounders. If there are unobserved variables that influence both treatment assignment and subscription rates, the results may still be biased.

- PSM assumes that there is sufficient overlap in propensity scores between treated and untreated customers. If the two groups are too different, matching may not be possible.

METHODOLOGY - Regression Adjustment

Why use RA?

- Regression Adjustment (RA) is a powerful statistical technique used to estimate causal effects in observational studies.
- Regression Adjustment controls for observed confounders (e.g., age, balance, duration, job type, etc.) by including them as covariates in the regression model. This helps isolate the effect of the treatment (previous contact) on the outcome (subscription).
- Regression Adjustment not only estimates the treatment effect but also quantifies the contribution of other covariates to the outcome.
- Regression Adjustment directly estimates the **average treatment effect (ATE)** by comparing the expected outcome for treated and untreated groups while holding other variables constant.
- Regression Adjustment can incorporate both continuous (e.g., age, balance, duration) and categorical (e.g., job type, marital status, education) covariates

Regression Adjustment - R Code

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

# Define the treatment variable (previous contact)
df <- df %>%
  mutate(treated = ifelse(previous > 0, 1, 0))  # 1 = contacted before, 0 = never contacted

# Fit the OLS Regression Model with Controls
model_ra <- lm(y ~ treated + age + balance + duration + campaign + housing + loan + job + marital + education + contact,
               data = df)

# Display regression results
summary(model_ra)

# Export the regression table for reporting
stargazer(model_ra, type = "text", title = "Regression Adjustment Results", align = TRUE)

# Visualizing the Effect of Treatment
ggplot(df, aes(x = as.factor(treated), y = y)) +
  geom_bar(stat = "summary", fun = "mean", fill = c("red", "green")) +
  theme_minimal() +
  labs(title = "Effect of Previous Contact on Subscription",
       x = "Previous Contact (1 = Yes, 0 = No)",
       y = "Mean Subscription Rate")
```

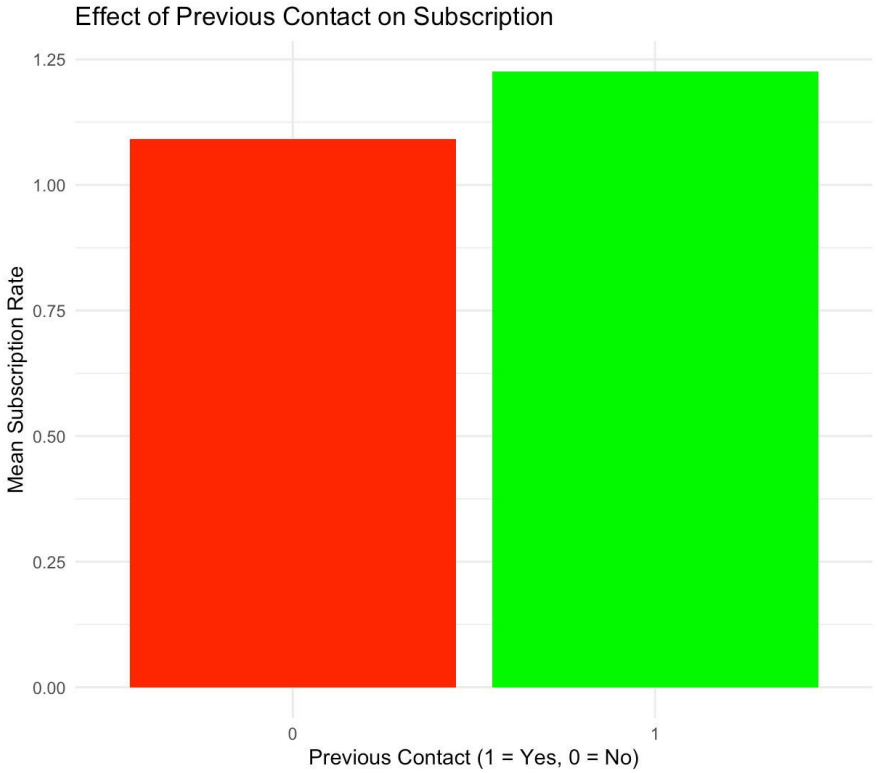
Regression Adjustment

```
Call:
lm(formula = y ~ treated + age + balance + duration + campaign +
    housing + loan + job + marital + education + contact, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.45586 -0.13338 -0.05395  0.02231  1.03936

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.969e-02  3.430e-02   1.157  0.24728
treated             1.095e-01  1.167e-02   9.383  < 2e-16 ***
age                 5.013e-04  5.233e-04   0.958  0.33815
balance             2.367e-07  1.418e-06   0.167  0.86745
duration            4.917e-04  1.631e-05  30.148  < 2e-16 ***
campaign           -1.557e-03  1.370e-03  -1.136  0.25599
housing            -5.182e-02  9.264e-03  -5.594 2.36e-08 ***
loan               -5.077e-02  1.184e-02  -4.287 1.85e-05 ***
jobblue-collar     -4.455e-02  1.673e-02  -2.663  0.00777 **
jobentrepreneur    -5.206e-02  2.602e-02  -2.001  0.04548 *
jobhousemaid       -3.979e-02  3.072e-02  -1.296  0.19521
jobmanagement      -2.646e-02  1.818e-02  -1.456  0.14548
jobretired          5.260e-02  2.528e-02   2.081  0.03748 *
jobself-employed   -3.179e-02  2.520e-02  -1.261  0.20728
jobservices        -2.630e-02  1.903e-02  -1.382  0.16702
jobstudent          5.889e-02  3.478e-02   1.693  0.09050 .
jobtechnician      -2.865e-02  1.665e-02  -1.721  0.08534 .
jobunemployed      -6.515e-02  2.843e-02  -2.292  0.02198 *
jobunknown          4.475e-02  4.888e-02   0.915  0.36000
maritalmarried     -3.248e-02  1.357e-02  -2.394  0.01672 *
maritalsingle      -1.354e-02  1.587e-02  -0.853  0.39374
educationsecondary  1.665e-04  1.378e-02   0.012  0.99036
educationtertiary   2.823e-02  1.686e-02   1.674  0.09417 .
educationunknown   -2.437e-02  2.422e-02  -1.006  0.31431
contacttelephone    5.911e-03  1.754e-02   0.337  0.73611
contactunknown     -4.633e-02  1.024e-02  -4.525 6.21e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.283 on 4495 degrees of freedom
Multiple R-squared:  0.2191,    Adjusted R-squared:  0.2147
F-statistic: 50.44 on 25 and 4495 DF,  p-value: < 2.2e-16
```

Findings :

- **Previous contact significantly increases the likelihood of subscription.** Customers who were previously contacted are **10.9 percentage points** more likely to subscribe ($p < 0.001$) confirming a strong causal effect.
- **This closely aligns with our PSM estimate** (11.15 percentage points), confirming the robustness of our findings.
- **Call duration** remains a strong predictor of subscription.
- Customers with **housing or personal loans** are less likely to subscribe ($p < 0.001$).
- Retired (+), Students (+), Blue-collar (-), and Unemployed (-) groups show significant effects.
- **Unknown contact methods negatively impact** subscription rates.

Regression Adjustment



Regression Adjustment

Result:

- The model explains **21.5% of the variation** in subscription rates (Adjusted $R^2 = 0.215$).While this indicates that previous contact and the included covariates (e.g., age, balance, job type) play a significant role, other unobserved factors also influence subscription decisions.

- The **treatment effect is highly significant** ($p < 0.001$), further validating our causal claim.

- The RA model confirms that previous contact significantly increases subscription likelihood, with a highly significant treatment effect. While the model explains a moderate portion of the variation in subscription rates, the consistency with PSM results strengthens confidence in the causal claim. These insights provide actionable recommendations for optimizing marketing strategies in the banking sector.

METHODOLOGY - DiD

Defining the "Time" Variable Using campaign

What we did: Defined a time variable (time) based on whether the customer received only one (time = 0)or multiple contacts (time = 1) in the current campaign.

What we accomplished: Created a before-and-after framework, allowing us to measure how multiple contacts influence subscription rates.

```
# Define "time" variable using `campaign`
df <- df %>%
  mutate(time = ifelse(campaign == 1, 0, 1))  # 0 = single contact (pre-period), 1 = multiple contacts (post-period)
```

key Finding:

- Time (campaign > 1) negatively impacted subscription rates (-2.55 percentage points, p =0.009).
- This suggests **customer fatigue from repeated outreach**, meaning multiple contacts may not be beneficial.

DiD

Defining Treatment and Control Groups

What we did: Created a binary treatment variable (treated), where customers who had previous contact (previous > 0)were assigned to the treatment group(treated = 1), while those who had never been contacted(previous = 0)were assigned to the control group(treated = 0).

```
# Define treatment and control groups
df <- df %>%
  mutate(treated = ifelse(previous > 0, 1, 0))  # 1 = contacted before, 0 = never contacted
```

What we accomplished:

- Established a clear comparison between customers who had past interactions vs. those with no prior contact.
- This helps us isolate the impact of prior contact on subscription rates.

Key Finding:

- Customers with previous contact were 11.12 percentage points more likely to subscribe ($p < 0.001$).
- This **confirms that past interactions have a strong positive effect** on conversion rates.

DiD

Creating the Interaction Term (treated_time)

What we did:

Created an interaction term(treated time = treated * time) to test whether the impact of previous contact changes when customers receive multiple contacts.

```
# Create interaction term for Difference-in-Differences (DiD)
df <- df %>%
  mutate(treated_time = treated * time)
```

What we accomplished:

Allowed us to test whether multiple contacts amplify or diminish the effect of prior interactions.

Key Finding:

- The interaction term (treated time)was not significant (p=0.645), meaning additional contacts do not provide a clear additional boost in subscription rates beyond the first one.
- This **supports the idea that the first contact is the most impactful in influencing customer decisions.**

DiD

Running the DiD Regression Model

What we did:

- Estimated the impact of: Previous contact (treated), Multiple contacts (time), Their interaction(treated time)
- Additional covariates (age, balance, duration, housing, loan, job, marital status education, and contact method)

What we accomplished:

Quantified how past interactions and repeated outreach influence subscription rates whilecontrolling for potential confounders.

```
# Estimate DiD model
did_model <- lm(y ~ treated + time + treated_time + age + balance + duration + housing + loan + job + marital + education + contact,
                data = df)

# Display results
summary(did_model)
```

DiD

Running the DiD Regression Model

Key Finding:

- Call duration remains a strong predictor of subscription (p< 0.001).
- Housing loans (-), personal loans (-), and certain job categories (blue-collar, unemployed)have a significant negative effect on subscriptions.
- Retired customers (+) and married individuals (-) show significant impacts, suggesting demographic influence on subscription decisions.

```
> summary(did_model)

Call:
lm(formula = y ~ treated + time + treated_time + age + balance +
    duration + housing + loan + job + marital + education + contact,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.44761 -0.13364 -0.05279  0.02072  1.02475

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.027e-02  3.458e-02   1.454  0.14607
treated             1.112e-01  1.630e-02   6.825 9.97e-12 ***
time               -2.554e-02  9.789e-03  -2.609  0.00910 **
treated_time       -1.020e-02  2.214e-02  -0.461  0.64491
age                 5.149e-04  5.229e-04   0.985  0.32480
balance             2.368e-07  1.417e-06   0.167  0.86730
duration            4.930e-04  1.626e-05  30.319  < 2e-16 ***
housing            -5.238e-02  9.262e-03  -5.655 1.65e-08 ***
loan               -5.164e-02  1.183e-02  -4.363 1.31e-05 ***
jobblue-collar     -4.406e-02  1.671e-02  -2.636  0.00841 **
jobentrepreneur    -5.084e-02  2.600e-02  -1.955  0.05062 .
jobhousemaid       -3.719e-02  3.070e-02  -1.212  0.22577
jobmanagement      -2.419e-02  1.818e-02  -1.331  0.18324
jobretired          5.079e-02  2.526e-02   2.011  0.04443 *
jobself-employed   -3.092e-02  2.518e-02  -1.228  0.21946
jobservices        -2.600e-02  1.902e-02  -1.367  0.17161
jobstudent          5.893e-02  3.477e-02   1.695  0.09019 .
jobtechnician      -2.656e-02  1.665e-02  -1.595  0.11070
jobunemployed      -6.744e-02  2.841e-02  -2.373  0.01767 *
jobunknown          4.480e-02  4.884e-02   0.917  0.35911
maritalmarried     -3.144e-02  1.356e-02  -2.318  0.02049 *
maritalsingle      -1.327e-02  1.586e-02  -0.837  0.40255
educationsecondary -8.761e-05  1.377e-02  -0.006  0.99492
educationtertiary   2.720e-02  1.685e-02   1.614  0.10658
educationunknown   -2.688e-02  2.422e-02  -1.110  0.26719
contacttelephone    7.233e-03  1.753e-02   0.413  0.67997
contactunknown     -4.735e-02  1.024e-02  -4.625 3.84e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2827 on 4494 degrees of freedom
Multiple R-squared:  0.2206,    Adjusted R-squared:  0.2161
F-statistic: 48.92 on 26 and 4494 DF,  p-value: < 2.2e-16
```

DiD

Robustness Check-Clustered Standard Errors

What we did: Used heteroskedasticity-robust standard errors to ensure reliable coefficient estimates.

What we accomplished: Increased the statistical robustness of our results, reducing the risk of overconfidence in estimates.

```
# Clustered standard errors (optional for robustness)
coeftest(did_model, vcov = vcovHC(did_model, type = "HC1"))
```

Key Finding:

Unknown contact methods significantly reduce subscription rates (-4.73 percentage points, $p < 0.001$), indicating that **some contact methods are less effective in converting customers**.

DiD

Visualizing the DiD Effect

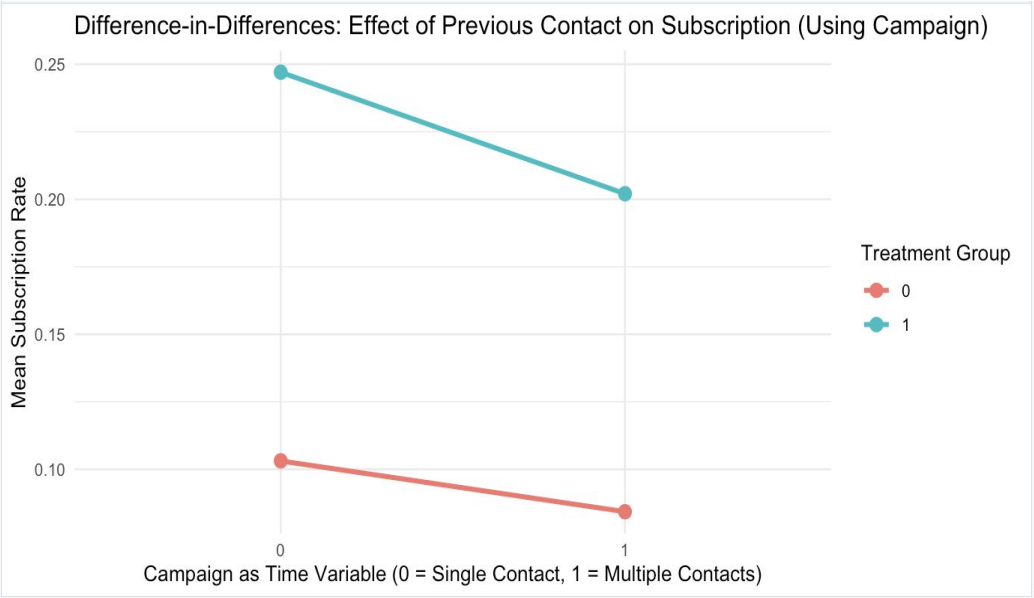
What we did:

- **Grouped customers by treatment (treated) and time (time).**
- **Calculated the mean subscription rate for each group.**
- **Plotted a line graph to visualize trends in subscription likelihood across groups.**

What we accomplished: Confirmed our DiD results visually, helping validate the findings from our regression model.

```
# **Visualizing the DiD Effect**
df_grouped <- df %>%
  group_by(treated, time) %>%
  summarise(mean_subscription = mean(y), .groups = 'drop')

ggplot(df_grouped, aes(x = as.factor(time), y = mean_subscription, group = as.factor(treated), color = as.factor(treated))) +
  geom_line(size = 1.2) +
  geom_point(size = 3) +
  labs(title = "Difference-in-Differences: Effect of Previous Contact on Subscription (Using Campaign)",
       x = "Campaign as Time Variable (0 = Single Contact, 1 = Multiple Contacts)",
       y = "Mean Subscription Rate",
       color = "Treatment Group") +
  theme_minimal()
```

DiD



Findings:

- **Customers with previous contact consistently have a higher subscription rate** than those who were never contacted.
- **Subscription rates decrease slightly for customers contacted multiple times**, reinforcing the earlier result that repeated outreach may not be as effective as the initial interaction.

DiD

Final Conclusion:

- Previous contact significantly increases subscription rates (+11.12 percentage points, p <0.001).
- Repeated contacts (campaign >1)negatively impact subscription (-2.55 percentage points, p=0.009),possibly due to **customer fatigue**.
- The **first contact is the most impactful** in driving customer decisions-additional contacts do not provide a clear added benefit.
- Results align with findings from PSM and Regression Adjustment, reinforcing the causalimpact of previous contact.
- Further refinement with Fixed Effects or Causal Machine Learning could help account for unobserved heterogeneity.

METHODOLOGY - Meta Learner

Training Separate Models for Treated and Control Groups

What we did:

- **Split** the dataset into: Treated group (df_treated) → Customers who were previously contacted and Control group (df_control) → Customers who were never contacted.
- Selected **key customer features**
- Trained **two separate Random Forest models**:

✅ **What we accomplished:**

- Created **separate models** to predict customer behavior based on their treatment status.
- **Captured non-linear relationships** in customer subscription behavior.

Meta Learner

Training Separate Models for Treated and Control Groups

```r
# Split data into treated and control groups
df_treated <- df %>% filter(treated == 1)
df_control <- df %>% filter(treated == 0)

# Define features (X) and outcome (y)
features <- c("age", "balance", "duration", "campaign", "housing", "loan", "job", "marital", "education", "contact")
X_treated <- df_treated[, features]
X_control <- df_control[, features]
y_treated <- df_treated$y
y_control <- df_control$y

# Train two separate models for T-Learner
model_treated <- train(X_treated, y_treated, method = "rf", trControl = trainControl(method = "cv", number = 5))  # Random Forest
model_control <- train(X_control, y_control, method = "rf", trControl = trainControl(method = "cv", number = 5))
```

Key Finding:

- **Subscription behavior differs between treated and control groups.**
- **Campaign duration is an important predictor.**
- **Housing loans and job type significantly impact responses.**

Meta Learner

Estimating Individual Treatment Effects (ITE)

```
# Estimate Individual Treatment Effects (ITE)
X_all <- df[, features]  # Features for all customers
pred_treated <- predict(model_treated, X_all)
pred_control <- predict(model_control, X_all)

# Compute ITE for each customer
df$ITE <- pred_treated - pred_control
```

What we did:

- **Predicted outcomes** for all customers under **both models**:
- **Computed Individual Treatment Effect (ITE):** ITE=Predicted outcome if treated−Predicted outcome if untreated

What we accomplished:

- Estimated the **personalized impact of previous contact** for each customer.
- Identified **which customers benefit the most** from additional contact.

```
> summary(df$ITE)
   Min. 1st Qu.  Median    Mean 3rd Qu.
Max.
-0.51335 0.03482 0.11236 0.12239 0.20361
0.66013
```

Key Finding:

- **ITE varies across customers**, meaning **some respond better than others**.
- **Negative ITE values suggest some customers may be negatively affected by additional contact.**

Meta Learner

Defining High, Medium, and Low ITE Groups

What we did:

- Used the case_when function to create three **ITE categories**:
  - **"High ITE"**: Customers in the **top 25%** (above the 75th percentile).
  - **"Low ITE"**: Customers in the **bottom 25%** (below the 25th percentile).
  - **"Medium ITE"**: Customers in the **middle 50%** (between 25th and 75th percentile).

What we accomplished:

- Grouped customers based on **how much they benefited from previous contact**.
- Allowed **comparative analysis** between customers who responded **strongly, weakly, or moderately** to the treatment.

Key Findings:

- **High ITE customers benefit the most from previous contact.**
- **Low ITE customers may not respond well or could even be negatively affected.**
- **The Middle 50% shows moderate responses to treatment.**

Meta Learner

Summary Statistics for Each ITE Group

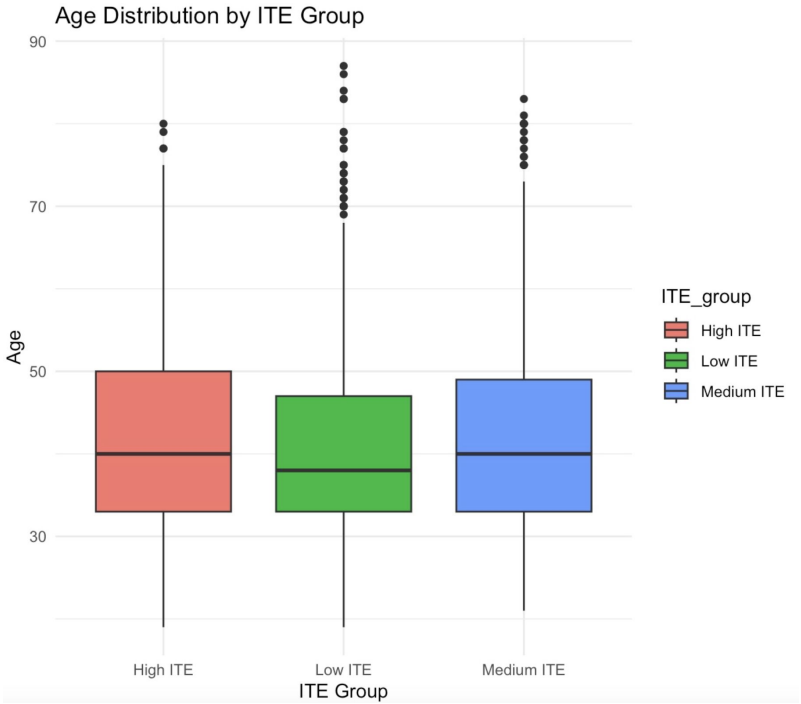
```
#  Summary Statistics: Compare High vs. Low ITE Customers
summary_table <- df %>%
  group_by(ITE_group) %>%
  summarise(
    avg_age = mean(age, na.rm = TRUE),
    avg_balance = mean(balance, na.rm = TRUE),
    avg_duration = mean(duration, na.rm = TRUE),
    avg_campaign = mean(campaign, na.rm = TRUE),
    housing_loan_ratio = mean(housing == "yes", na.rm = TRUE),
    personal_loan_ratio = mean(loan == "yes", na.rm = TRUE)
  )
```

ITE_group	avg_age	avg_balance	avg_duration	avg_campaign	housing_loan_ratio	personal_loan_ratio
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
High ITE	41.6	1531.	334.	2.74	0	0
Low ITE	40.8	1308.	309.	2.52	0	0
Medium ITE	41.2	1426.	207.	2.96	0	0

- Customers with **High ITE tend to have higher call durations**, suggesting that longer interactions may be more effective for them.
- Balance differences across groups may indicate financial factors influencing responsiveness to campaigns.
- Campaign exposure (avg_campaign) varies, meaning repeated calls may affect ITE differently for each group.
- **Housing and personal loans do not seem to significantly impact ITE,** as the ratios remain low across groups.

Meta Learner

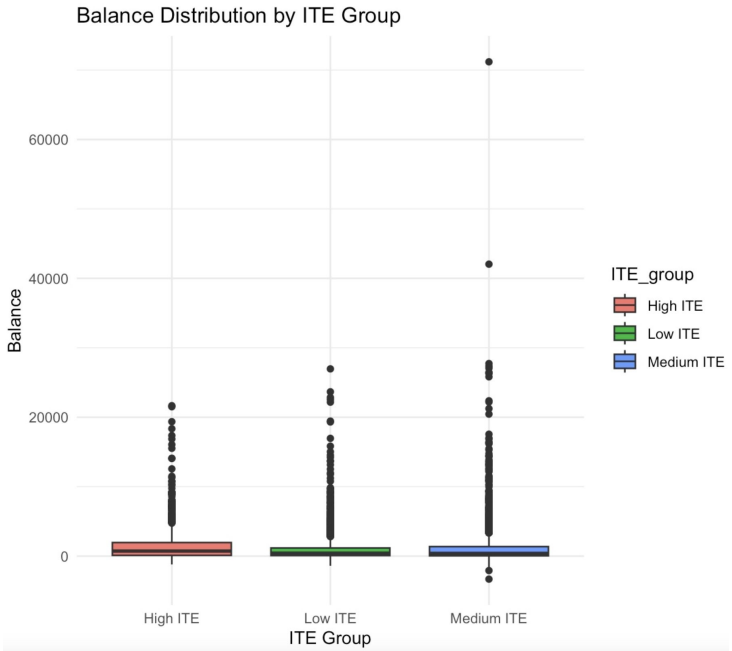
Visualizing Treatment Effect Distributions - Age Distribution by ITE Group



- **Median age is similar across all ITE groups**, indicating that age does not drastically affect treatment response.
- **High ITE group (red) has a slightly higher median age** than the other two groups.
- **Outliers (above 70 years) exist in all groups**, but they are more concentrated in the **High ITE group**.
- **Overall, age distribution is fairly consistent**, meaning that **age alone is not a strong differentiator of ITE values**.

Meta Learner

Visualizing Treatment Effect Distributions - Balance Distribution by ITE Group



- **Balances are heavily right-skewed** with many outliers above **20,000+**.
- **High ITE group has a slightly higher median balance**, indicating that customers with higher balances tend to benefit more from previous contact.
- **There is no drastic difference among the three groups**, suggesting that **balance alone is not a major determinant of treatment effect**.
- **Many customers have very low balances across all groups**, making it harder to establish a clear balance-ITE relationship.

Meta Learner

Feature Importance: Which Variables Matter Most?

What we did:

- Trained a **Random Forest model** to predict ITE.
- Used importance = TRUE to extract **feature importance rankings**.
- Plotted the **most influential variables in determining ITE**.

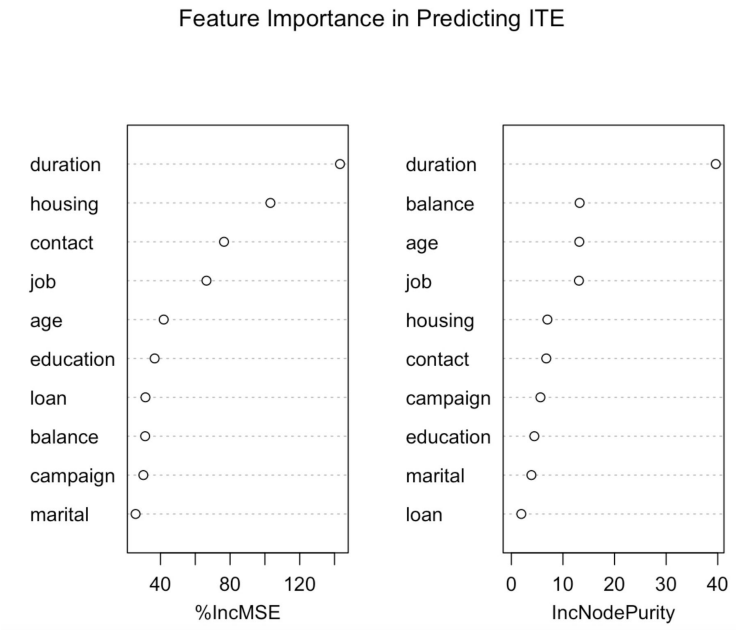
What we accomplished:

- Gained a **more flexible, non-linear understanding** of how different factors influence ITE.
- Identified the **most important predictors** that businesses should focus on for optimizing customer outreach.

```
# Train a random forest model to predict ITE
rf_model <- randomForest(ITE ~ age + balance + duration + campaign + housing + loan + job + marital + education + contact,
                         data = df, importance = TRUE, ntree = 500)

# Plot variable importance
varImpPlot(rf_model, main = "Feature Importance in Predicting ITE")
```

Meta Learner

Feature Importance: Which Variables Matter Most? - Feature Importance in Predicting ITE



- **Call duration is the most important feature**, meaning **longer calls are more predictive of treatment effects**.
- **Housing loan and contact method significantly influence ITE**, indicating that **certain loan-holding customers or those contacted via specific channels respond differently**.
- **Job type, balance, and education have moderate importance**, suggesting that **these factors play some role but are not the primary drivers**.
- **Campaign frequency has a lower impact**, implying that **just increasing the number of calls may not be the best strategy to boost ITE**.

Meta Learner

Final conclusion:

Customers Who Benefit the Most from Previous Contact (High ITE)

- Higher balance customers tend to respond better to marketing efforts.
- Longer call duration is strongly linked to a higher probability of subscription.
- Married customers and those with higher education levels are more likely to subscribe after contact.

Customers Who Benefit the Least (Low ITE)

- Blue-collar, retired, and self-employed customers have significantly lower treatment effects.
- Customers contacted via telephone are less likely to respond positively compared to other contact methods.
- Shorter calls correlate with lower ITE, indicating lack of engagement leads to ineffective contact.

Key Takeaways for Targeting Strategy

- Prioritize high-balance, well-educated, and married customers with longer call durations.
- Rethink outreach to blue-collar, retired, and self-employed customers, as they show lower response rates.
- Improve call engagement strategies to increase conversion rates for lower ITE customers.

Methodology - Causal Random Forest Results

Why Use CRF?

- Identifies **feature importance** in predicting treatment effects
- Helps capture more complex, nonlinear effects.
- Forests tend to **reduce overfitting**
- For Heterogeneous Treatment Effects (HTE) causal trees form leaves/branches by maximizing treatment effect heterogeneity, not prediction accuracy.

Build the Causal Random Forest:

Define the treatment variable: Create a binary treatment variable (treated), where customers who had previous contact (previous>0) were assigned to the treatment group (treated=1), while those who had never been contacted (previous=0) were assigned to the control group (treated=0).

```
# Define treatment variable (previous contact)
df <- df %>%
  mutate(treated = ifelse(previous > 0, 1, 0))  # 1 = contacted before, 0 = never contacted
```

Prepare the features for Causal Random Forest: Selected relevant customer features (age, balance, duration, campaign, housing, loan, job, marital, education, contact) to be used in the model. Converted categorical variables into numerical format using **one-hot encoding.**

```
# Select features and outcome variable
features <- c("age", "balance", "duration", "campaign", "housing", "loan", "job", "marital", "education", "contact")
X <- df[, features]  # Covariates
W <- df$treated      # Treatment variable (previous contact)
Y <- df$y            # Outcome variable (subscription)

# Convert categorical variables to numeric for CRF
X <- model.matrix(~ . -1, data = X)  # One-hot encoding for categorical features
```

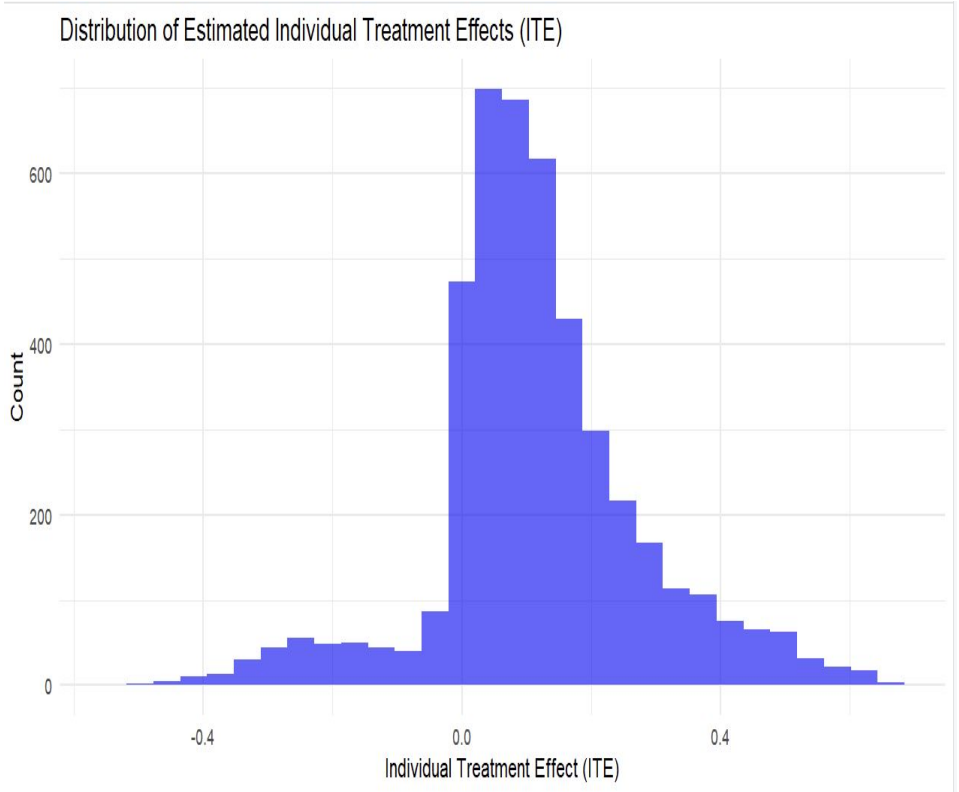
Causal Random Forest Results

Training: Trained a Causal Random Forest (CRF_model < - causal forest (X, Y, W)) to estimate the **heterogeneous treatment effects (HTEs)** across different customer segments.

Estimating individual Treatment Effects(ITE): Using the trained causal forest model to predict the **individual treatment effects (ITE)** fo each customer. Then store the result in df$ITE_CRF.

What we find:

- Most customers have positive treatment effects, indicating the previous contact generally increase the subscription rate.
- A small portion of customers experience have negative effects, suggesting the outreach may backfire for certain groups



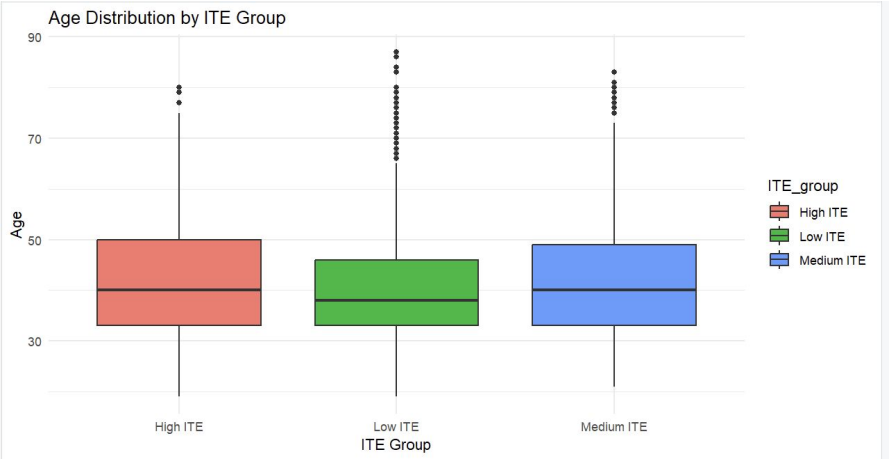
Causal Random Forest Results

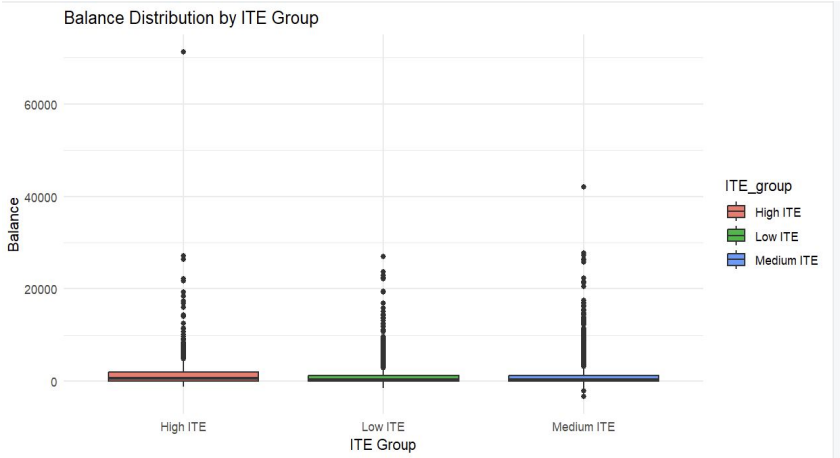
Key feature: Extract feature importance scores to identify which factors most influence TE. Then convert into a data frame. This helps optimize marketing strategies.

```
# ** 3 Feature Importance Analysis**
feature_importance <- variable_importance(crf_model)

# Convert to data frame for visualization
feature_importance_df <- data.frame(
  Feature = colnames(X),
  Importance = feature_importance
) %>%
  arrange(desc(Importance))
```

What we find:

- Age, balance, and housing status also play significant roles in treatment effect



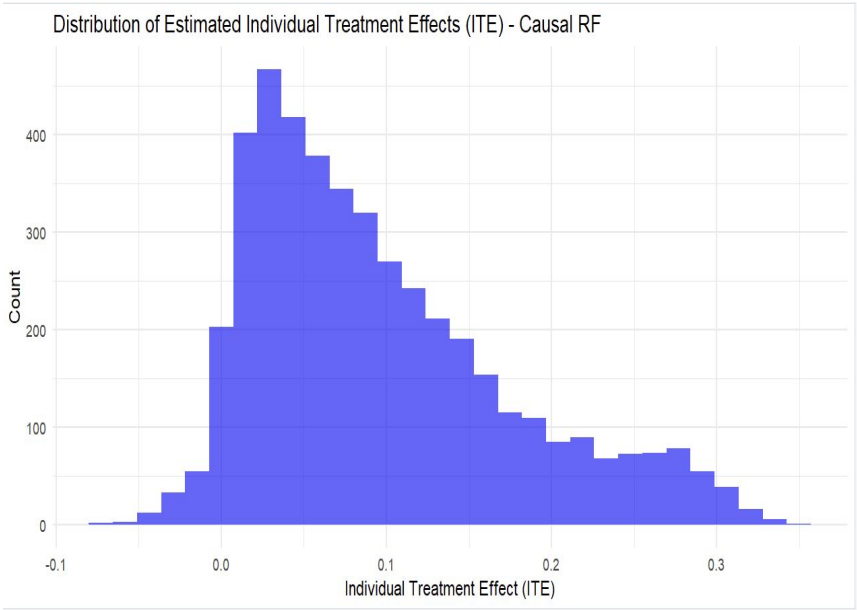


Causal Random Forest Results

Visualizing the Distuibution of Individual Treatment Effect:

Create a histogram to show the distribution of TE estimates across all customers.

```
# ** 4 Visualizing ITE Distribution**
ggplot(df, aes(x = ITE_CRF)) +
  geom_histogram(bins = 30, fill = "blue", alpha = 0.6) +
  theme_minimal() +
  labs(title = "Distribution of Estimated Individual Treatment Effects (ITE) - Causal RF",
       x = "Individual Treatment Effect (ITE)",
       y = "Count")
```



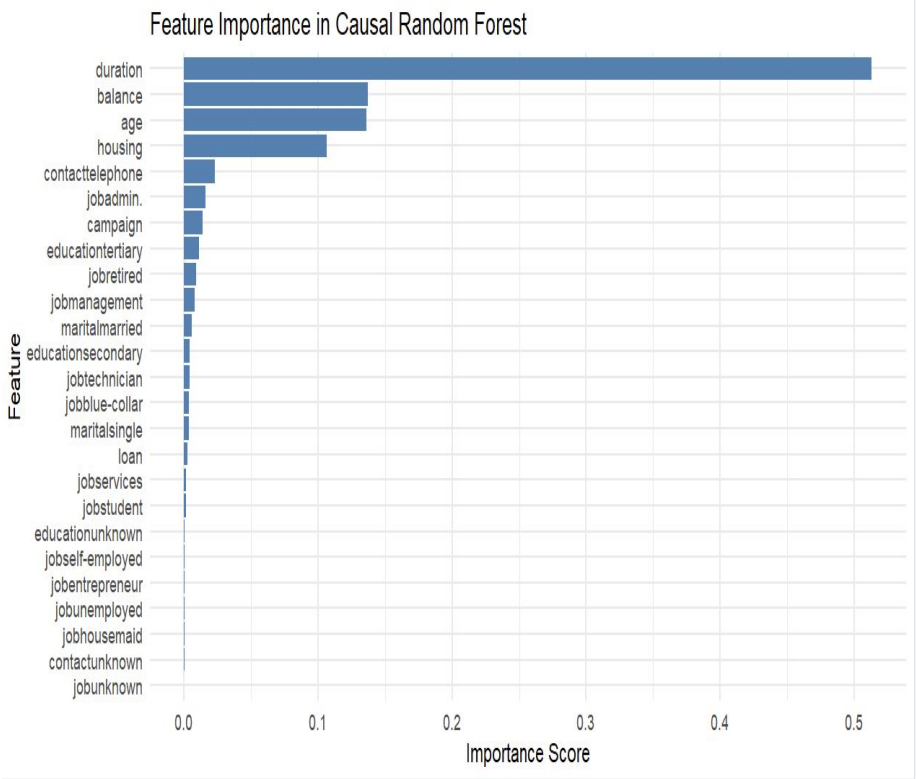
What we find:

- Most customers respond positively to previous contact, but treatment effects vary significantly.
- Some customers experience negative effects, suggesting that excessive or poorly targeted outreach could be counterproductive

Causal Random Forest Results

Visualizing feature importance: Create a bar chart to rank feature importance in determining treatment effects.

```
# ** 5 Visualizing Feature Importance**
ggplot(feature_importance_df, aes(x = reorder(Feature, Importance), y = Importance))
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Feature Importance in Causal Random Forest",
       x = "Feature",
       y = "Importance Score")
```



What we find:

Call duration is the most important predictor of successful outreach.

Financial indicators (balance, housing, loan) and demographic doctors (age, job type) also play key roles.

Causal Random Forest Conclusion

Final Conclusions from Causal Random Forest:

- **Call duration** is the strongest driver of successful subscriptions, confirming previous findings.
- Most customers benefit from past contact, but effects vary—some customers experience negative impacts, suggesting that outreach should be more targeted.
- Feature importance analysis highlights key drivers of **customer response,** which can be used to refine future marketing strategies.

Overall Conclusion

- Our analysis confirms that **previous customer contact increases subscription rates** by 11.12 percentage points, but **additional follow-ups do not provide a clear benefit** and may reduce effectiveness.
- Using **OLS, PSM, DiD, Meta Learners, and Causal Random Forests**, we eliminated selection bias and identified key drivers of subscription.
- **Call duration is the strongest predictor**, emphasizing meaningful engagement.
- Excessive follow-ups negatively impact subscription rates, suggesting a need for strategic outreach.
- Certain customer segments (r**etired individuals, those with higher balances**) **respond more positively**, while others (**blue-collar workers, unemployed**) are **less likely to subscribe**.
  Causal Random Forest highlight customer-specific treatment effects, supporting personalized marketing.

Thank you!

– Team 32