

Predicting Customer Responses in Telemarketing

Part 1:

Report Overview:

This report analyzes customer behaviour in a Portuguese bank's telemarketing campaigns with the goal of identify the key factors that drive acceptance of the bank's long-term deposit offers. By identifying and investigating key customer attributes, this report will provide actionable insights to refine telemarketing strategies, improve customer conversion rates and optimize resource allocation.

Exploration of the Dataset:

We first investigate the key variables to understand what influences customer decisions in these telemarketing campaigns. We explore how age, call duration, occupation, and previous campaign outcomes impact the likelihood of accepting long-term deposits. This will allow the bank to further focus on identified high-acceptance groups and refine their approach to lower acceptance ones.

Investigation of the Dataset's Variables:

The following examines the distribution of the dataset’s categorical variables and basic statistics for the numerical ones. This will highlight the imbalance in customer responses and the impact of factors like call duration, age, and occupation, which will be explored further in subsequent sections. **Table 1.1** shows the distribution of the categorical variables and **Table 1.2**, the numerical ones.

R code for Table 1.1

```
12 #Looking at the distinct values of the categorical attributes.
13 table(mdata$job)
14 table(mdata$marital)
15 table(mdata$education)
16 table(mdata$default)
17 table(mdata$housing)
18 table(mdata$loan)
19 table(mdata$contact)
20 table(mdata$month)
21 table(mdata$day_of_week)
22 table(mdata$days)
23 table(mdata$previous)
24 table(mdata$spoutcome)
25 table(mdata$y)
```

Table 1.1

[illegible]

(Table 1.1) reveals six variables with unknown values (highlighted in red) that need to be removed before further analysis. The 'y' variable reveals a significant imbalance in the dataset, with only 11% of customers accepting the offer. This imbalance demonstrates the need for improvement in identifying customers likely to accept long-term deposits. All customers were contacted by phone, so call duration will be a factor that requires consideration.

Statistics of the Numerical Variables:

The following examines the dataset's numerical variables as well as presents several statistics. Special focus will be given to customer age and call duration due to their influence on campaign outcome that will be expanded upon in further analyses. **(Table 1.2)** shows the max, min, median, and mean of applicable variables.

R code for Table 1.2

```
394 #looks at the numeric variables from the dataset
395 numeric_data <- mdata %>%
396   select_if(is.numeric)
397
398 summary(numeric_data)
```

Table 1.2

Numeric Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
age	18	32	38	40.12	47	88
duration	0	103	181	256.8	317	3643
campaign	1	1	2	2.539	3	35
pdays	0	999	999	960.2	999	999
previous	0	0	0	0.1907	0	6
emp.var.rate	-3.4	-1.8	1.1	0.085	1.4	1.4
cons.price.idx	92.2	93.08	93.75	93.58	93.99	94.77
cons.conf.idx	-50.8	-42.7	-41.8	-40.5	-36.4	-26.9
euribor3m	0.635	1.334	4.857	3.621	4.961	5.045
nr.employed	4964	5099	5191	5166	5228	5228

(Table 1.2) reveals that most customers fall within the age range of 32 to 47, indicating that targeting younger or older segments may present opportunities for expanding the customer base. The average call duration is around five minutes, suggesting that customers make decisions quickly, and longer calls may indicate increased or decreased consideration of the offer. Further analysis will determine its impact.

Investigation of Customer Attributes

We now closely examine selected customer attributes, including age, occupation, and the outcome of previous telemarketing campaigns. They were chosen based on their influence on the outcome of the call and will be demonstrated in the following visualizations.

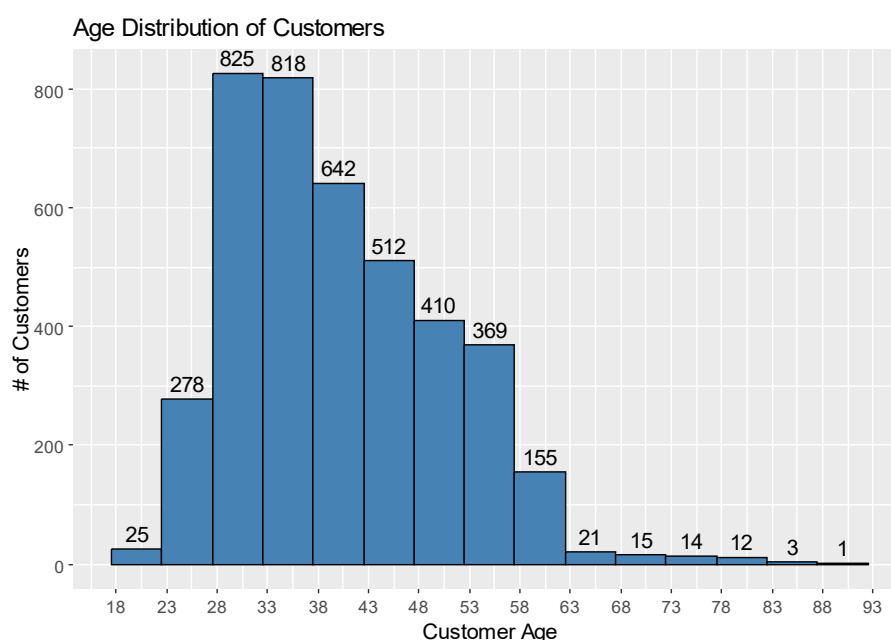
Age Distribution of Customers:

Through analysing customer age, we identify potential areas for increasing the customer base, particularly among younger age groups. **(Figure 1.1)** shows the distribution of the bank's customer ages.

R code for Figure 1.1

```
467
468 #distribution of the customer's ages
469 ggplot(mdata, mapping = aes(x = age)) +
470   geom_histogram(binwidth = 5, color = "black", fill = "steelblue") +
471   scale_x_continuous(breaks = seq(min(mdata$age), max(mdata$age), by = 5)) +
472   labs(title = "Age Distribution of Customers", x = "Customer Age", y = "# of Customers")+
473   stat_bin(binwidth = 5, geom = "text", aes(label = ..count..), vjust = -0.5) # Adds count labels on top of bars
474
```

Figure 1.1



(Figure 1.1) shows that **87%** of customers are aged 28 to 58, likely more financially stable and potentially interested in long-term deposits. The lack of customers under 28 and over 58 indicates an opportunity to increase the customer base. The bank could introduce retirement planning seminars for customers approaching 58, while offering personalized savings plans for younger customers in the 28-35 age range. This will potentially create more recipients for future campaigns.

Customer Occupation and Age:

Here, we conduct a deeper analysis of the bank's customers' occupations and their respective age distributions. **(Figure 2.1)** illustrates the job distribution and **(Figure 2.2)**, the corresponding age patterns.

R code for Figure 2.1

```
373 #Looking at the distribution of customers jobs|
374 job_data <- mdata %>%
375   filter(job != 'unknown') %>%
376   group_by(job) %>%
377   summarise(num_customers = n())
378
379 ggplot(data = job_data, mapping = aes(x = reorder(job, num_customers), y = num_customers, fill = job))+
380   geom_bar(stat = 'identity')+
381   labs(title = "Distribution of Customer's jobs", x = "Jobs", y = "# of Customers")+
382   geom_text(aes(label = num_customers))+
383   coord_flip()+
384   theme(legend.position = 'none')
385
```

R code for Figure 2.2

```
399 #Looking at the average age of customers by jobs
400 job_data <- mdata %>%
401   filter(job != 'unknown') %>%
402   mutate(job = fct_rev(fct_infreq(job)))
403
404 ggplot(data = job_data, mapping = aes(x = job, y = age, fill = job))+
405   geom_boxplot()+
406   coord_flip()+
407   labs(title = 'Distribution of Customer Age by Job', x = 'Customer Jobs', y = 'Customer Age')+
408   theme(legend.position = 'none')+
409   scale_y_continuous(breaks = seq(0,90, by = 5))
410
```

Figure 2.1

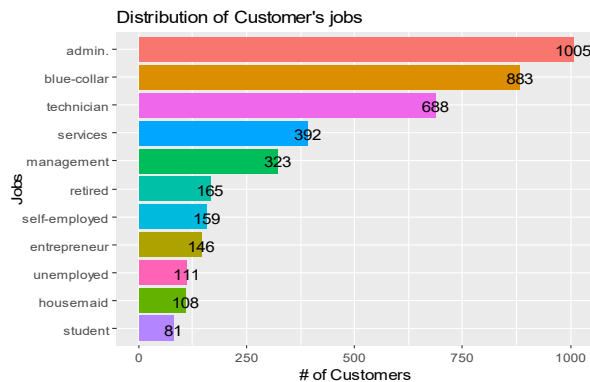
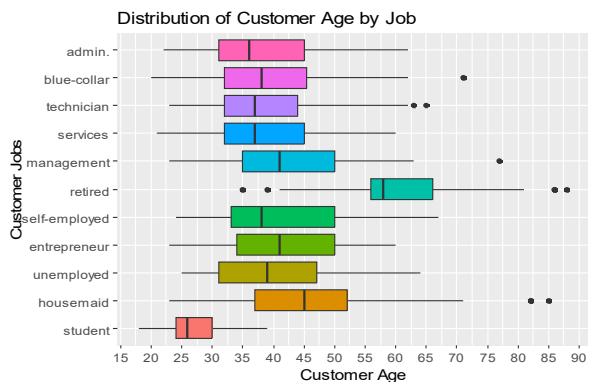


Figure 2.2



A large portion of the bank's customers fall into administrative, blue-collar, and technician roles **(Figure 2.1)**, with most aged 35 to 45 **(Figure 2.2)**. In contrast, students, housemaids, and the unemployed are smaller segments **(Figure 2.1)**, likely due to lower incomes. Age patterns align with expectations, students aged **25-30** and retirees **55-65**. The bank could offer tailored financial products such as income protection plans, and retirement planning workshops to better engage these smaller subsets.

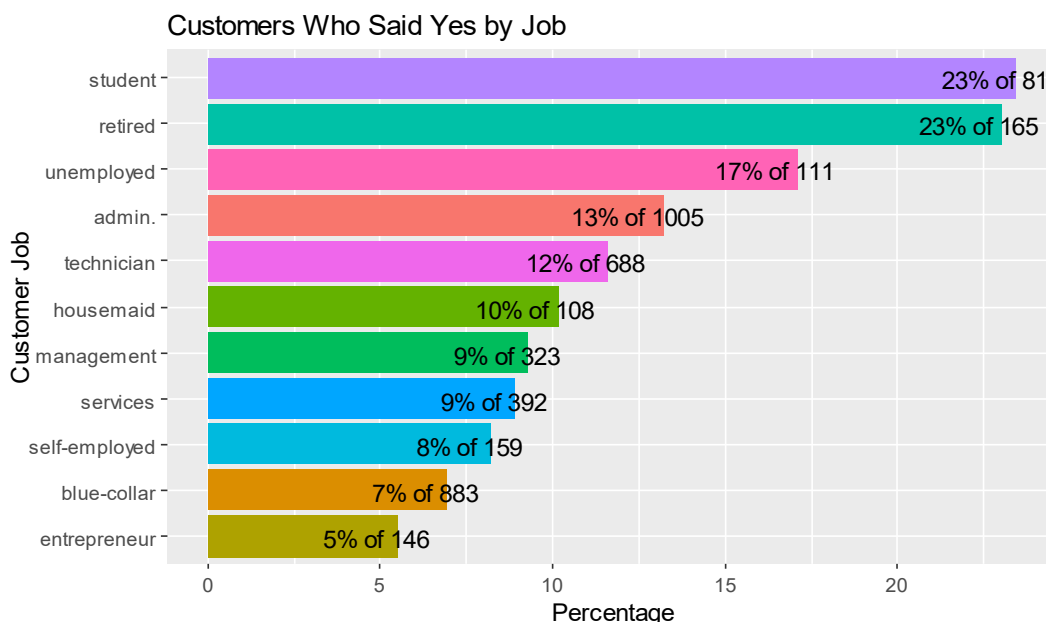
Customer Occupation and Its Impact:

This section examines the readiness of customers to accept the long-term deposit based on their job categories. **(Figure 3.1)** reveals that students, retirees, and unemployed had the highest rates of acceptance.

R code for Figure 3.1

```
427 #Percentage of Customers who said 'yes' by job
428 pjob_data <- mdata %>%
429   filter(job != 'unknown') %>%
430   group_by(job) %>%
431   summarize(total = n(), yes_customers = sum(y == 'yes')) %>%
432   mutate(percent_yes = (yes_customers / total) * 100)
433
434 ggplot(pjob_data, mapping = aes(x = reorder(job, percent_yes), y = percent_yes, fill = job))
435   geom_bar(stat = 'identity')+
436   coord_flip()+
437   labs(title = "Customers Who Said Yes by Job", x = 'Customer Job', y = 'Percentage')+
438   geom_text(aes(label = paste0(round(percent_yes, 0), "% of ", total)), hjust = 0.7)+
439   theme(legend.position = 'none')
```

Figure 3.1



(Figure 3.1) shows that students and retirees, despite being smaller segments, had the highest acceptance rates at 23%, likely due to retirees seeking financial security and students aiming to start saving early. Unemployed customers followed with a 17% acceptance rate, highlighting the appeal of financial stability during unemployment. In contrast, entrepreneurs had the lowest acceptance rate at 5%, likely due to a need for financial flexibility. Administrative workers, blue-collar workers, and technicians, though making up a significant portion of the customer base, had lower acceptance rates (13%, 12%, and 5%, respectively). Offers could be made to target these important, yet low-acceptance subsets such as career advancement loans and packages that allow for flexible cash flow in order to raise their acceptance rate.

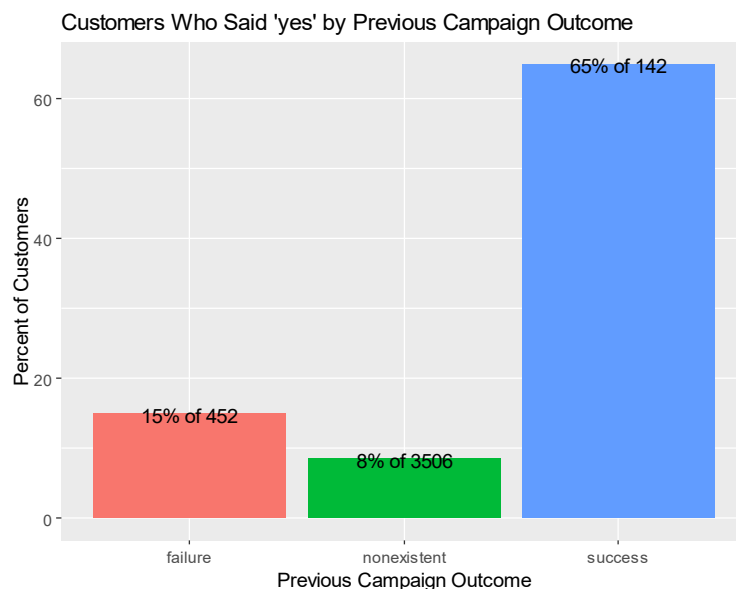
Influence of Previous Campaigns on Future ones:

The outcome of previous campaigns significantly influences customer acceptance in current telemarketing efforts. **(Figure 4.1)** shows us the considerably positive impact a previously successful campaign has on future campaigns. We also see the detrimental impact a failed, or non-existent campaign has.

R code for Figure 4.1

```
483 #looking at the percentage of customers who said yes based on their previous campaign outcome
484 c_data <- mdata %>%
485   group_by(poutcome) %>%
486   summarize(total = n(), yes_customers = sum(y == 'yes')) %>%
487   mutate(percent_yes = (yes_customers / total) * 100)
488
489 ggplot(c_data, mapping = aes(x = poutcome, y = percent_yes, fill = poutcome))+
490   geom_bar(stat = 'identity')+
491   geom_text(aes(label = paste0(round(percent_yes,0), "% of ", total))) +
492   labs(title = "Customers who Said 'yes' by Previous Campaign Outcome",
493        x = "Previous Campaign Outcome", |
494        y = "Percent of Customers")+
495   theme(legend.position = 'none')
```

Figure 4.1



As shown in **(Figure 4.1)**, a previously successful campaign strongly influences customer acceptance, with **65%** of those customers again saying 'yes'. Positive interactions boost the likelihood of future success while even a failed campaign still had **15%** of customers still accepted the offer. This is still more than those customers with no previous campaign attempt at **8%**. The data shows even previous contact, regardless of outcome, makes customers more open to future offers. This indicates a blanket increase in campaigns will still increase customer acceptance rates, even if only for future ones.

Part 1 - Conclusion:

With just **11%** of customers accepting the offer, improved targeting is needed, both for larger job segments like administrative workers and technicians and to smaller, yet high-acceptance segments like students, retirees, and the unemployed. A specific focus should be on students as targeted campaigns at students could initiate a trickledown effect as they join the workforce increasing acceptance rates among the larger subsets they eventually join. This is further substantiated by the displayed impact the presence a previous campaign has on customer acceptance rates, successful or otherwise.

Part 2:

Model Reasoning:

In the interest of predicting campaign outcomes, a logistic regression model was used with 'job' and call 'duration' as the primary inputs. Customer occupation had a demonstrated effect on customer acceptance rates and call duration reflects the customer's interest during the interaction. As the goal is to predict a binary outcome ('yes' or 'no'), a logistic regression model is the natural choice.

Model Development:

Prior to model development, the dataset was split into training (**80%**) and testing (**20%**) data to prevent overfitting and bias. Below, we see the R code used to split the data, train the model and then its summary statistics.

R code for splitting data set and logistical model

```

560
561 #cleaning the dataset before putting it in the model
562 clean_data <- mdata %>%
563   filter(job != 'unknown')
564
565 #splitting the dataset into training and testing datasets and making y a factor
566 split = sample.split(clean_data$y, SplitRatio = 0.8)
567
568 training_data <- subset(clean_data, split == TRUE)
569 training_data$y <- as.factor(training_data$y)
570
571 testing_data <- subset(clean_data, split == FALSE)
572 testing_data$y <- as.factor(testing_data$y)
573
574 #creating the logistical Model
575 log_mod_job_duration <- glm(y ~ job * duration, data = training_data, family = 'binomial')
576 summary(log_mod_job_duration)

```

Summary of logistical model

```

Console Terminal Background Jobs
R 4.4.1 ~ /

Call:
glm(formula = y ~ job * duration, family = "binomial", data = training_data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.2070177   0.1948698  -16.457 < 2e-16 ***
jobblue-collar -1.6388857   0.4061253   -4.035 5.45e-05 ***
jobentrepreneur -1.0932277   0.7731181   -1.414 0.15735
jobhousemaid   -0.8490573   0.8339789   -1.018 0.30864
jobmanagement  -0.4723814   0.4451097   -1.061 0.28857
jobretired      1.2378381   0.3882720    3.188 0.00143 **
jobself-employed -0.2717010   0.5857611   -0.464 0.64276
jobservices     -0.1754963   0.3902184   -0.450 0.65290
jobstudent      1.1652703   0.5126360    2.273 0.02302 *
jobtechnician   -0.1543684   0.3185520   -0.485 0.62796
jobunemployed    0.7238865   0.5263777    1.375 0.16906
duration         0.0039381   0.0004123    9.551 < 2e-16 ***
jobblue-collar:duration 0.0016262   0.0007480    2.174 0.02971 *
jobentrepreneur:duration -0.0003588   0.0011713   -0.306 0.75938
jobhousemaid:duration 0.0018050   0.0017290    1.044 0.29650
jobmanagement:duration 0.0002328   0.0009187    0.253 0.79998
jobretired:duration -0.0017071   0.0008644   -1.975 0.04829 *
jobself-employed:duration -0.0008397   0.0009941   -0.845 0.39828
jobservices:duration -0.0004372   0.0008408   -0.520 0.60309
jobstudent:duration -0.0018395   0.0011286   -1.630 0.10313
jobtechnician:duration 0.0002998   0.0007061    0.425 0.67119
jobunemployed:duration -0.0008310   0.0012497   -0.665 0.50606
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2254.2  on 3248  degrees of freedom
Residual deviance: 1773.3  on 3227  degrees of freedom
AIC: 1817.3

Number of Fisher Scoring iterations: 6

```

The model's summary indicates that students and retirees are significantly more likely to accept the long-term deposit offer, while blue-collar workers and self-employed individuals are less likely, aligning with earlier findings. Additionally, each additional second in call durations boost the chances of acceptance by about 0.4%, making it a key factor in telemarketing success. However, the effect of call duration differs by job; it positively influences blue-collar workers but slightly decreases acceptance for retirees.

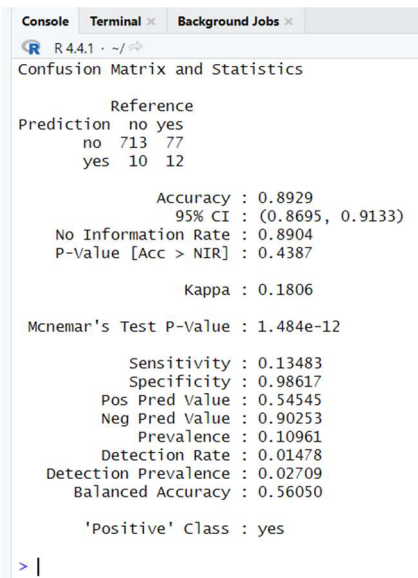
Model Performance:

The initial performance of the model shows a high, yet misleading accuracy score and ultimately proves to be of little use with the default threshold.

R code for creating a confusion matrix for the model

```
578 #Creating a table of predictions for the testing_data
579 prediction_numbers <- predict(log_mod_job_duration, testing_data, type = 'response')
580
581 #Converting likelihood numbers into outcomes
582 prediction_outcomes <- ifelse(prediction_numbers > 0.5, 'yes', 'no')
583 prediction_outcomes <- as.factor(prediction_outcomes)
584
585 #Creating a confusion matrix to compare predictions with outcomes
586 con_matrix <- confusionMatrix(prediction_outcomes, testing_data$y, positive = 'yes')
587 (con_matrix)
```

Confusion matrix and statistics of the model



```
Console Terminal Background Jobs
R 4.4.1 ~ /
Confusion Matrix and Statistics

      Reference
Prediction no yes
no      713  77
yes      10  12

      Accuracy : 0.8929
      95% CI   : (0.8695, 0.9133)
      No Information Rate : 0.8904
      P-Value [Acc > NIR] : 0.4387

      Kappa : 0.1806

      Mcnemar's Test P-Value : 1.484e-12

      Sensitivity : 0.13483
      Specificity : 0.98617
      Pos Pred Value : 0.54545
      Neg Pred Value : 0.90253
      Prevalence : 0.10961
      Detection Rate : 0.01478
      Detection Prevalence : 0.02709
      Balanced Accuracy : 0.56050

      'Positive' Class : yes

> |
```

The model's predictions yielded 12 true positives, 713 true negatives, 77 false negatives, and 10 false positives. Despite an overall accuracy of **89%**, this figure is misleading due to the imbalance in the dataset, where most responses were 'no.' The No Information Rate (**NIR**) is also **89%**, indicating the model doesn't outperform simply guessing 'no' for all cases.

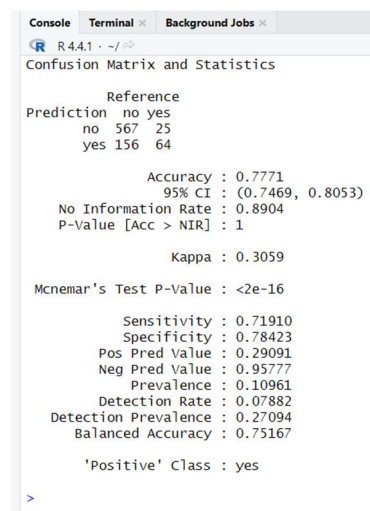
Tuning the Model

The model shows improvement when the threshold for predicting 'yes' is lowered from 0.5 to 0.1. The default threshold led to a majority of 'yes' responses being missed; reducing it capture more positive cases and improves the model's overall usefulness.

R code for creating a confusion matrix for the model (threshold set: 0.1)

```
578 #Creating a table of predictions for the testing_data
579 prediction_numbers <- predict(log_mod_job_duration, testing_data, type = 'response')
580
581 #Converting likelihood numbers into outcomes
582 prediction_outcomes <- ifelse(prediction_numbers > 0.1, 'yes', 'no')
583 prediction_outcomes <- as.factor(prediction_outcomes)
584
585 #Creating a confusion matrix to compare predictions with outcomes
586 con_matrix <- confusionMatrix(prediction_outcomes, testing_data$y, positive = 'yes')
587 (con_matrix)
```

Confusion matrix and statistics of the model (threshold set: 0.1)



```
Console Terminal Background Jobs
R 4.4.1 ~ /
Confusion Matrix and Statistics

      Reference
Prediction no yes
no      567  25
yes     156  64

      Accuracy : 0.7771
      95% CI   : (0.7469, 0.8053)
      No Information Rate : 0.8904
      P-Value [Acc > NIR] : 1

      Kappa : 0.3059

      Mcnemar's Test P-Value : <2e-16

      Sensitivity : 0.71910
      Specificity : 0.78423
      Pos Pred Value : 0.29091
      Neg Pred Value : 0.95777
      Prevalence : 0.10961
      Detection Rate : 0.07882
      Detection Prevalence : 0.27094
      Balanced Accuracy : 0.75167

      'Positive' Class : yes
>
```

lowering the threshold improved the model's predictions in identifying 'yes' responses, with 64 true positives, 567 true negatives, 25 false negatives, and 156 false positives. Although the overall accuracy dropped from **89% to 78%**, the model's performance became more balanced. Sensitivity increased significantly to **72%**, showing a much better ability to identify 'yes' customers, compared to the previous sensitivity of **13%**. However, this came at the cost of reduced specificity, which fell from **99% to 78%**.

Part 2 - Conclusion:

Lowering the prediction threshold improved the model's ability to identify customers likely to accept long-term deposit offers at the expense of more false positives. However, the bank's interests will always be ensuring high-potential customers are not missed. This strategy is necessary in a business where each successful conversion significantly contributes to revenue growth. By focusing on customers identified as likely to say 'yes,' the bank can prioritize follow-up campaigns and allocate marketing resources efficiently, ensuring higher conversion rates and improved return on investments.

Bibliography

Equitable Equations (2023) *Logistic Regression in R* [Online Video] Available at: <https://youtu.be/E7J3M1oYVlc?feature=shared> Accessed: [15/09/2024]

Grolemund, G., & Wickham, H. (2017). *R for Data Science*. O'Reilly Media.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *An Introduction to Statistical Learning: With Applications in R*. New York: Springer.

R Programming 101 (2022) *Linear Regression using R Programming* [Online Video] <https://www.youtube.com/watch?v=-mGXnm0fHtI> Accessed [09/09/2024]

StatQuest with Josh Statrmer (2018) *Logistic Regression in R, Clearly Explained!!!!* [Online Video] Available at: https://youtu.be/C4N3_XJJ-jU?feature=shared Accessed: [17/09/2024]

University of Essex, (2024) *Principles of Data Visualization 2 Lecture cast*. Available at: https://www.my-course.co.uk/mod/scorm/player.php?a=15144¤torg=articulate_rise&scoid=30353&sesskey=RmcJwnEFPK&display=popup&mode=normal Accessed: [01/09/2024]