DA 6813 Case Study 1

AUTHOR
Will Hyltin, Holly Milazzo and Tim Harrison

PUBLISHED
September 22, 2024

1 Executive Summary

This study aimed to develop a predictive model to determine whether clients of a Portuguese banking institution would subscribe to a term deposit following a marketing campaign. The analysis focused on a dataset containing client demographics, financial indicators, and marketing campaign details. Key variables such as age, contact type, month of the campaign, number of previous contacts, employment rate, and previous client contacts were found to influence the likelihood of subscription.

Due to the unbalanced nature of our response variable, a balanced sample was taken so the model would not favor prediction for those who did not subscribe versus those that did. Then, a logistic regression model was built using backward stepwise selection based on the Akaike Information Criterion (AIC). Initially, the model had high specificity (86.36%), meaning it effectively identified clients that did not subscribe. However, despite balancing train and test sets, it underperformed in identifying clients who did subscribe, as reflected by its lower sensitivity (56.99%).

To further improve the model's overall performance, an optimal decision threshold was identified, balancing sensitivity and specificity. After optimization, the model achieved a balanced accuracy of 73.48%, with both sensitivity and specificity improving to 73.12% and 73.86%, respectively. This adjustment enhanced the model's ability to accurately classify both subscribing and non-subscribing clients.

The results provide actionable insights for the bank, enabling it to more effectively target clients and optimize marketing resources. The findings suggest that further improvements, such as testing alternative models or incorporating more detailed features, could yield even higher predictive performance.

2 Problem Statement

The task of this case study is to develop a predictive model to classify whether clients of a Portuguese banking institution will subscribe to a term deposit following a direct marketing campaign. The campaigns were based on phone calls, and often, multiple contacts were made to the same client to assess if they would subscribe ('yes') or not ('no'). The goal is to accurately predict the likelihood of a client subscribing to a term deposit based on various input variables, including demographic, social, and economic indicators, as well as information from previous marketing campaigns.

Key variables influencing the prediction include the client's age, job type, marital status, education, and financial status (e.g., default history, housing loan, and personal loan). Additionally, variables related to the marketing campaign, such as the type of contact, day, and month of the last contact, and previous campaign outcomes, are also critical in determining the client's response. Social and economic context attributes, such as the employment variation rate and consumer confidence index, are included to enhance the model's predictive power.

The primary objective is to build a classification model that can accurately predict whether a client will subscribe to a term deposit. This will allow the bank to target its marketing efforts more effectively, optimizing resource allocation, and improving conversion rates.

3 Methodology

The analysis began with data preparation, where the dataset of 21 variables, both categorical and numeric, was cleaned and transformed. Categorical variables were converted to factors for proper handling in models, and the target variable, `y`, was transformed into a binary indicator, where 1 indicated a client subscribed to a term deposit and 0 indicated otherwise. The `pdays` variable, which represented the number of days since a previous contact, was recoded into a binary indicator named `pcontact` to simplify the analysis.

Exploratory data analysis (EDA) was conducted to explore the distribution and relationships within the data. Box plots were generated to visualize the distribution of numeric variables based on the target outcome, and bar plots were used to explore the frequency distribution of categorical variables. Additionally, a correlation matrix was constructed to examine the relationships among numeric variables and to identify potential multicollinearity issues.

The dataset was then split into a training set (80%) and a test set (20%) to ensure model evaluation was conducted on unseen data. To address class imbalance, as there were significantly more clients who did not subscribe to the term deposit, resampling techniques were applied to create a balanced dataset for training. This was done by narrowing down to all responses where a bank customer subscribed and an equal number of those who did not subscribe before performing the train and test split.

Logistic regression was selected as the primary modeling technique. A stepwise backward selection method, based on the Akaike Information Criterion (AIC), was used to remove insignificant variables and select the most parsimonious model. To mitigate multicollinearity, variables with high variance inflation factor (VIF) values were removed.

The model was evaluated using metrics such as accuracy, sensitivity, and specificity, and the results were summarized in a confusion matrix. Given that class imbalance persisted, even after balancing the data, an optimal cutoff threshold for classifying clients was determined by balancing sensitivity and specificity. This threshold optimization helped improve the performance of the logistic regression model by identifying the point where sensitivity and specificity converged.

4 Data

The dataset contains 4,119 rows and 21 variables. The variables are split between 11 categorical (character type) and 10 numeric variables. In preparation for analysis, categorical variables were converted into factors to ensure proper handling in models. The target variable, y, was converted into a binary outcome where "yes" was mapped to 1, indicating that the client subscribed to a term deposit, and "no" was mapped to 0.

Name	df3
Number of rows	4119
Number of columns	21
________________________	
Column type frequency:	
factor	10
numeric	11
_________________________	

Group variables	None

Data summary

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
job	0	1	FALSE	12	adm: 1012, blu: 884, tec: 691, ser: 393
marital	0	1	FALSE	4	mar: 2509, sin: 1153, div: 446, unk: 11
education	0	1	FALSE	8	uni: 1264, hig: 921, bas: 574, pro: 535
default	0	1	FALSE	3	no: 3315, unk: 803, yes: 1
housing	0	1	FALSE	3	yes: 2175, no: 1839, unk: 105
loan	0	1	FALSE	3	no: 3349, yes: 665, unk: 105
contact	0	1	FALSE	2	cel: 2652, tel: 1467
month	0	1	FALSE	10	may: 1378, jul: 711, aug: 636, jun: 530
day_of_week	0	1	FALSE	5	thu: 860, mon: 855, tue: 841, wed: 795
poutcome	0	1	FALSE	3	non: 3523, fai: 454, suc: 142

Variable type: numeric

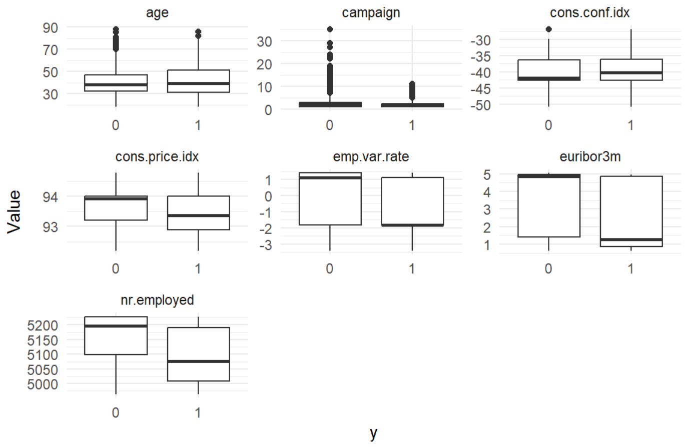
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	40.11	10.31	18.00	32.00	38.00	47.00	88.00	▅▇▅▁▁
duration	0	1	256.79	254.70	0.00	103.00	181.00	317.00	3643.00	▇▁▁▁▁
campaign	0	1	2.54	2.57	1.00	1.00	2.00	3.00	35.00	▇▁▁▁▁
previous	0	1	0.19	0.54	0.00	0.00	0.00	0.00	6.00	▇▁▁▁▁
emp.var.rate	0	1	0.08	1.56	-3.40	-1.80	1.10	1.40	1.40	▁▂▁▁▇
cons.price.idx	0	1	93.58	0.58	92.20	93.08	93.75	93.99	94.77	▁▆▃▇▂
cons.conf.idx	0	1	-40.50	4.59	-50.80	-42.70	-41.80	-36.40	-26.90	▅▇▁▇▁
euribor3m	0	1	3.62	1.73	0.64	1.33	4.86	4.96	5.04	▅▁▁▁▇
nr.employed	0	1	5166.48	73.67	4963.60	5099.10	5191.00	5228.10	5228.10	▁▁▃▁▇
y	0	1	0.11	0.31	0.00	0.00	0.00	0.00	1.00	▇▁▁▁▁
pcontact	0	1	0.04	0.19	0.00	0.00	0.00	0.00	1.00	▇▁▁▁▁

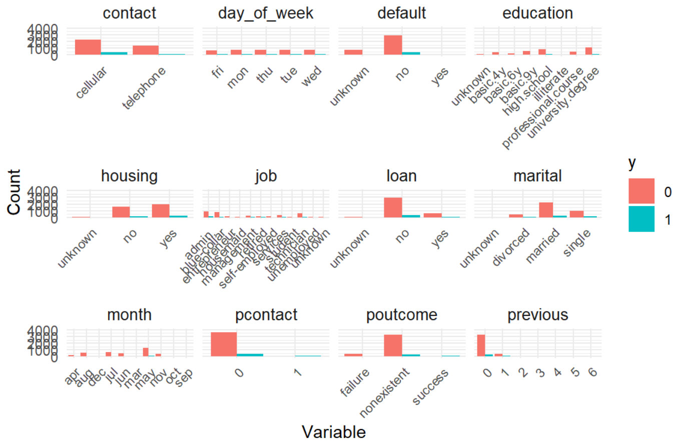
During the exploratory data analysis (EDA), various visualizations were utilized to examine the distributions and relationships within the dataset. Box plots were used to display the distribution of numeric variables such as **age**, **campaign**, **emp.var.rate**, **cons.price.idx**, **cons.conf.idx**, **euribor3m**, and **nr.employed**, stratified by whether or not the client subscribed to a deposit. These visualizations helped identify differences in the distributions between clients who subscribed and those who did not.

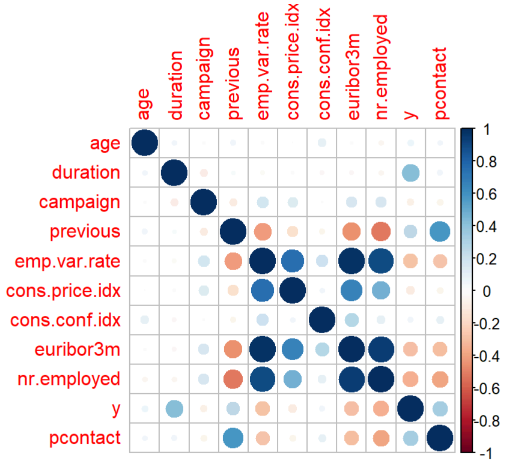
Bar plots were also used to explore the distribution of categorical variables like **job**, **marital status**, **education**, **default**, **housing**, **loan**, **contact**, **month**, **day_of_week**, and **poutcome**. These plots illustrated the relative frequencies of each category concerning the target outcome, shedding light on which factors might be more indicative of a client's decision to subscribe.

Regarding the **pdays** variable, a value of 999 indicated that a client had not been previously contacted, while any other number signified prior contact. To simplify the analysis, this variable was transformed into a binary indicator, where "1" denotes previous contact and "0" indicates no prior contact.

Finally, to explore relationships among numeric variables, a correlation matrix was generated. This matrix was visualized using a heatmap to provide a clear representation of the strength and direction of correlations between variables.







5 Findings

The logistic regression model was built using a backward stepwise elimination process based on the Akaike Information Criterion (AIC). During each step, insignificant variables such as job, day_of_week, education, marital status, and others were removed to improve model parsimony and fit. The final model included age, contact, month, campaign, nr.employed, and pcontact as the key predictors.

Two sets of results were generated to evaluate the performance of the model—one based on the initial logistic regression with a non-optimal threshold and the other with an optimized threshold for better classification performance.

For the **non-optimal threshold**, the confusion matrix showed an accuracy of 71.27%, with a specificity of 86.36% and a sensitivity of 56.99%. The model had higher specificity, meaning it was better at correctly identifying clients who did not subscribe, but it performed less well in identifying those who did.

After **threshold optimization**, the model's performance improved. The confusion matrix for the optimized model yielded an accuracy of 73.48%, with more balanced specificity (73.86%) and sensitivity (73.12%). This optimization process allowed for better classification of both subscribing and non-subscribing clients by adjusting the decision threshold to balance the trade-off between specificity. and sensitivity

Overall, the findings indicate that the threshold adjustment improved the model's ability to predict both positive and negative outcomes more evenly. This highlights the importance of optimizing decision thresholds, particularly in imbalanced datasets like this one. Interestingly, this issue was still encountered despite balancing the initial audience data based on the response value, suggesting that an imbalance in the predictors may also influence a binary classification model's performance.

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	674	744.9079	838.9079
- job	11	10.1992377	685	755.1072	827.1072
- day_of_week	4	0.9350717	689	756.0423	820.0423
- education	6	5.8197873	695	761.8620	813.8620
- marital	3	1.8152111	698	763.6772	809.6772
- cons.price.idx	1	0.0286753	699	763.7059	807.7059
- default	1	0.0548683	700	763.7608	805.7608
- cons.conf.idx	1	0.2981325	701	764.0589	804.0589
- housing	2	2.4672400	703	766.5262	802.5262
- previous	1	1.0908448	704	767.6170	801.6170
- poutcome	2	3.4459295	706	771.0629	801.0629

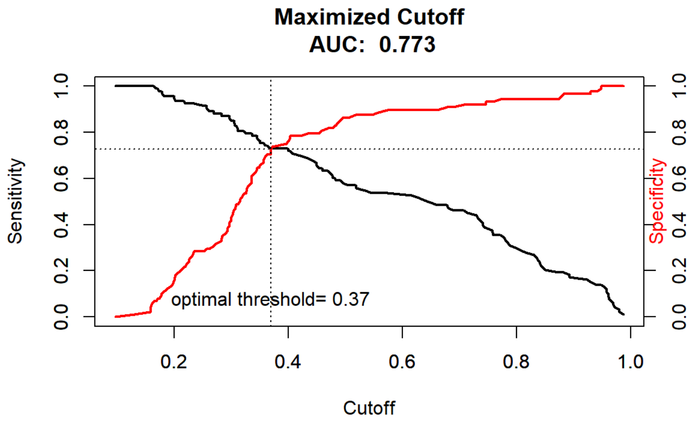
AIC at Each Step

	0	1
0	76	40
1	12	53

Confusion Matrix Non-Optimal Value

Metric	Value
Accuracy	0.7127
Sensitivity	0.5699
Specificity	0.8636

Model Results Non-Optimal Value



	0	1
0	65	25
1	23	68

Confusion Matrix Optimal Value

Metric	Value
Accuracy	0.7348
Sensitivity	0.7312
Specificity	0.7386

Model Results Optimal Value

```
as.factor(y) ~ age + contact + month + campaign + nr.employed +
    pcontact
```

6 Conclusion

The analysis successfully developed a logistic regression model to predict whether clients would subscribe to a term deposit following a marketing campaign. Through data preparation, exploratory analysis, and model building, key predictors such as age, contact, month, campaign, employment status (nr.employed), and previous contact (pcontact) were identified as significant factors influencing client subscription decisions.

The initial model had a higher specificity but lower sensitivity, meaning it was more effective at identifying clients who did not subscribe but less so for those who did. After optimizing the decision threshold, the model achieved a more balanced performance, with improved accuracy and sensitivity. This highlights the value of threshold adjustment in classification problems, particularly when dealing with imbalanced datasets.

In conclusion, the model offers actionable insights for the bank to more effectively target potential customers, allowing for better allocation of marketing resources and improved conversion rates. Further enhancements, such as testing additional models or incorporating more complex feature engineering, could further improve predictive performance.

7 Appendix

```
pacman::p_load(MASS, tidyverse, here, car, corrplot, skimr, caret, ggplot2, dplyr, tidyr, ROCR,knitr)
df1 <- read.csv(here('Case Study 1','bank-additional.csv'), sep = ';')
df2 <- df1 %>% mutate_if(is.character, as.factor)
df3 <- df2 %>% mutate(
  pcontact = ifelse(pdays == 999, 0, 1),
  y = ifelse(y == 'yes',1,0)
) %>% select(!pdays)

df_long <- df3 %>%
  pivot_longer(cols = c(age, campaign, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed),
               names_to = "variable", values_to = "value")
```

```r
ggplot(df_long, aes(x = as.factor(y), y = value)) +
  geom_boxplot() +
  facet_wrap(~variable, scales = "free") +
  labs(x = "y", y = "Value") +
  theme_minimal()

dflong2 <- df3 %>% mutate(
  previous = as.factor(previous),
  pcontact = as.factor(pcontact)
)

df_long2 <- dflong2 %>%
  pivot_longer(cols = c(job, marital, education, default, housing, loan, contact, month, day_of_week, poutcome, previous, pcontact),
               names_to = "variable", values_to = "value")

ggplot(df_long2, aes(fill = as.factor(y), x = value)) +
  geom_bar(position = "dodge") +
  facet_wrap(~variable, scales = "free_x", nrow = 3) +
  labs(x = "Variable", y = "Count", fill = "y") +
  ylim(0, 4000) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 8),
    strip.text = element_text(size = 10)
  )

cor1 <- df3 %>% select_if(is.numeric) %>% cor()
corrplot(cor1)

df4 <- df3 %>% dplyr::select(!duration)
set.seed(321)
trn_part <- sample(nrow(df4),0.8*nrow(df4),replace = F)
dftrain <- df4[trn_part,]
dftest <- df4[-trn_part,]

df_sub = df4 %>% filter(y == 1)
df_no_sub = df4 %>% filter(y == 0)
sample_no_sub = sample_n(df_no_sub, nrow(df_sub))
df_bal = rbind(sample_no_sub,df_sub)

set.seed(321)
tr_ind_bal <- sample(nrow(df_bal), 0.8 * nrow(df_bal), replace = FALSE)
dftrain_bal <- df_bal[tr_ind_bal, ]
dftest_bal <- df_bal[-tr_ind_bal, ]

m7.log <- step(glm(as.factor(y) ~ . - loan - emp.var.rate - euribor3m,
                   data = dftrain_bal, family = binomial),
               direction = "backward", trace = 0)

predprob_final <- predict(m7.log, newdata = dftest_bal, type = "response")
pr_class_final <- ifelse(predprob_final > 0.5, 1, 0)

predprob2_log_bal <- predict(m7.log, newdata = dftest_bal, type = "response")
predclass2_log_bal <- ifelse(predprob2_log_bal >= 0.5, 1, 0)

conf_matrix <- caret::confusionMatrix(as.factor(pr_class_final), as.factor(dftest_bal$y))
cm1_table <- as.data.frame.matrix(conf_matrix$table)
accuracy <- round(conf_matrix$overall['Accuracy'], 4)
sensitivity <- round(conf_matrix$byClass['Sensitivity'], 4)
specificity <- round(conf_matrix$byClass['Specificity'], 4)

pred_bal <- prediction(predprob2_log_bal,dftest_bal$y)
auc_bal <- round(as.numeric(performance(pred_bal, measure = "auc")@y.values),3)

plot(unlist(performance(pred_bal, "sens")@x.values), unlist(performance(pred_bal, "sens")@y.values),
     type="l", lwd=2,
     ylab="Sensitivity", xlab="Cutoff", main = paste("Maximized Cutoff\n","AUC: ",auc_bal))

par(new=TRUE)

plot(unlist(performance(pred_bal, "spec")@x.values), unlist(performance(pred_bal, "spec")@y.values),
     type="l", lwd=2, col='red', ylab="", xlab="")
axis(4, at=seq(0,1,0.2))
mtext("Specificity",side=4, col='red')

min.diff_bal <- which.min(abs(unlist(performance(pred_bal, "sens")@y.values) - unlist(performance(pred_bal, "spec")@y.values)))
min.x_bal<-unlist(performance(pred_bal, "sens")@x.values)[min.diff_bal]
min.y_bal<-unlist(performance(pred_bal, "spec")@y.values)[min.diff_bal]
optimal_bal <-min.x_bal

abline(h = min.y_bal, lty = 3)
abline(v = min.x_bal, lty = 3)
text(min.x_bal,0,paste("optimal threshold=",round(optimal_bal,2)), pos = 3)

pr_class_bal = ifelse(predprob2_log_bal > optimal_bal, 1, 0)
```

```r
conf_matrix <- caret::confusionMatrix(as.factor(pr_class_bal), as.factor(dftest_bal$y))
cm_table <- as.data.frame.matrix(conf_matrix$table)
accuracy <- conf_matrix$overall['Accuracy']
sensitivity <- conf_matrix$byClass['Sensitivity']
specificity <- conf_matrix$byClass['Specificity']

results_df <- data.frame(
  Value = c(round(accuracy, 4), round(sensitivity, 4), round(specificity, 4))
)

kable(cm_table, format = "pipe", align = "c", caption = "Confusion Matrix Optimal Value")
kable(results_df, format = "pipe", align = "c", col.names = c("Metric", "Value"), caption = "Model Results Optimal Value")
final_formula <- formula(m7.log)
print(final_formula)
```