Marketing A Term Deposit

Chenyu Zhou

Summary

This article explores factors affecting people's purchase decision of a term deposit during a bank's marketing campaign. Using a hierarchical logistic model, I found that consumers who have purchased a term deposit in previous marketing campaigns are more likely to purchase again, but the effect of previous campaign outcomes diminishes when consumers are more confident about future economy.

1. Introduction

A term deposit offered at banks is a type of investment associated with relatively low risks. Consumers' deposits are locked up for a period of time. At the end of the investment period, consumers withdraw their deposits and gain some interests in return. Compared to traditional saving accounts, term deposits offer a slightly higher interest rate. With consumers' term deposits, banks could make a profit by lending these money to other individuals or companies in need and charging an even higher interest rate. Therefore, banks always try to run marketing campaigns and attract more term deposits.

With a Portuguese bank marketing dataset, my primary research goal is to identify crucial features of a successful telephone marketing campaign for term deposits. Specifically, I wonder how consumers' decisions in previous marketing campaigns affect the current campaign outcome. Moro, Cortez, and Rita (2014) [1] explored a similar data set with four different data mining models. Focusing on model comparisons, they found neural network is the best model to predict the success of banks' telephone marketing, compared to logistic regression, decision trees, and support vector machine. Different from their approach, in additional to my primary research goal, I also examined if the odds of purchasing a term deposit differ across contacted month with a hierarchical logistic model.

2. Data

The dataset is based on marketing campaigns of a Portuguese bank from May 2008 to November 2010. This dataset is available in the UCI machine learning repository. Due to data limitation, the current data set only includes a part of the dataset used by Moro, Cortex and Rita (2014). Overall, 41,188 observations of 18 variables are included in the dataset.

The binary response variable represents whether consumers purchased a term deposit at the end of a marketing campaign. All potential predictors are grouped from four aspects and listed in Table 1. Indicated by recent contact type, consumers are contacted via telephone or cellular. Three potential outcomes are recorded for previous campaigns. Success indicates

[1] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems (2014).

Table 1: Variables

Group	Variables
Consumer Characteristics	age (num), job (factor), marital status (factor), education (factor), housing loan history (factor), personal loan history (factor)
Recent Contact	type (factor), month (factor), day (factor)
Contact History	number of contacts performed with this consumer for this campaign (num), number of days since approached by previous marketing campaigns (num), number of previous campaigns the consumer was approached (num), previous campaign outcome (factor)
Social & Economic	quarterly employment variation rate (num),daily Euribor 3 month rate (num), monthly consumer price index (num), monthly consumer confidence index (num)

that a consumer has purchased a term deposit in previous marketing campaigns, while failure indicates that a consumer has been approached before, but has never purchased a term deposit. Nonexistent indicates a new consumer who has never been contacted for a marketing campaign.

Four social and economic attributes are listed. The quarterly employment variation rate describes the economy from the aspect of employment. The consumer price index suggests the price variation of goods and services. The consumer confidence index shows consumer optimism of the current and future household spending and saving activities. The Euribor 3 month rate provides the average interest rate that European banks charge each other for loans with three-month maturity, indicating banks' expectation of future economic conditions. A low consumer price index, a high consumer confidence index, and a low Euribor 3 month rate indicate that people are optimistic about future economy.

Missing data is a potential problem in this data set. Less than 10 percent of data for education, housing loan history, personal loan history, and consumers' job title is missing. I used MICE in R to create imputations and take care of this problem. Five imputed datasets were generated and these imputed datasets follow a similar distribution as the observed dataset. The following analysis is based on the first randomly imputed dataset.

3.Exploratory Analysis

Plotting potential predictors and the response variable, I observed a nonlinear relationship of age. Specifically, consumers are less likely to purchase the product during their mid-ages (30s-50s), compared to their 20s and 60s (Figure 1). For further analysis, I regrouped consumers' age into three categories: young (age<=30), middle (30<age<60), and old (age>=60). Marital status, housing loan, personal loan, and contacted day are similar across consumers who purchased a term deposit and those who did not. Consumers are more likely to purchase a term deposit if they have high school degrees, and are contacted by cellular in March, September, October, and December (Table 8, 11, and 12). Since the likelihood of purchasing a term deposit varies across contacted month, a varying-intercept model needs to be considered. Additionally, the number of contacts made before this campaign is

positively related with the outcome of this campaign, while the number of contacts made for this campaign is negatively related with the likelihood of purchasing a term deposit (Figure 4 and 2). Furthermore, consumers who have previously purchased a term deposits are more likely to purchase again than those who have never been approached before or those who have been approached but have not purchased a term deposit in previous marketing campaigns (Table 2).

Table 2: Previous Campaign Outcome and Term Deposits

	failure	nonexistent	success
no	0.86	0.91	0.35
yes	0.14	0.09	0.65

Next, I examined how social and economics attributes affect consumers' decision of purchasing a term deposit during a marketing campaign. Since the Euribor 3 month rate is clustered around 1.3 and 4.9, I regrouped the variable into two categories: low (euribor<=3) and high (euribor>3). Plotting social and economics attributes against the response variable, a negative relationship is observed for consumer price index and Euribor 3 month rate, while a positive relationship is observed for consumer confidence index. The pattern is consistent with their definitions. Additionally, I found that the relationship between consumer confidence index and consumers' probability of purchasing a term deposit varies across previous campaign outcome (Figure 10). The significance of this interactions will be analyzed in the model below.

4. Modeling

In this section, logistic models and multi-level models are used to explore features leading to a successful marketing campaign for term deposits. Most numeric variables, including age, quarterly employment variation rate, monthly consumer price index, and monthly consumer confidence index are mean-centered for precise interpretation. The model selection process is summarized in Table 4.

M**1**: Baseline Model

Based on results from exploratory analysis and the main research question, all potential predictors are included in the baseline model. As indicated in the exploratory analysis section, I included new categorical variables of age and Euribor 3 month rate instead of the numerical ones to avoid violations of the linearity assumption. Additionally, I dropped employment variation rate due to multicollinearity.

M**2**: Interactions

Since I am also interested in if the effect of previous campaign outcome on the probability of consumers purchasing a term deposit varies by expected economy, the interaction term between previous campaign outcome and consumer confidence index is added to the advanced model.

M**3**: Stepwise Model

According to stepwise model selection results, AIC drops variables including housing loan and personal loan. Since the anova test shows that the new model is not significantly different from the interaction model, I continued with the stepwise model for simplicity.

M**4**: Hierarchical Logistic model

To examine if the odds of purchasing a term deposit differ across contacted month, a varying intercept by contacted month is added to the stepwise model. The following multilevel logistic model is my final model:

$$\begin{aligned}\log(\frac{\pi_{ij}}{1-\pi_{ij}}) &= \beta_0 + \gamma_{0j} + \beta_1 \text{newage}_{ij} + \beta_2 \text{job}_{ij} + \beta_3 \text{marital}_{ij} + \beta_4 \text{education}_{ij} \\ &+ \beta_5 \text{contact}_{ij} + \beta_6 \text{day}_{ij} + \beta_7 \text{campaign}_{ij} + \beta_8 \text{previous}_{ij} + \beta_9 \text{poutcome}_{ij} \\ &+ \beta_{10} \text{confc}_{ij} + \beta_{11} \text{pricec}_{ij} + \beta_{12} \text{loweuri}_{ij} + \beta_{13} \text{poutcome}_{ij} * \text{confc}_{ij}, \\ &\epsilon_{ij} \sim N(0, \sigma^2), \gamma_{0j} \sim N(0, \gamma_0^2)\end{aligned}$$

. Model assumptions are checked with binned residual plots. Most data points are randomly distributed within the 95 percent band (Figure 11). The pattern in the plot between binned residual against numerical variables also looks random, indicating no violation for the linearity assumption (Figure 13). In the logistic model (M3), all VIF values are below 10, indicating no serious multicollinearity problem in the final model.

Visualizing the random effect, March and May are potential outliers (Figure 14). Their 95 percent confidence intervals are far away from zero. The odds of purchasing a term deposit is extremely high in March, but low in May. After eliminating outliers, the model is computed again again. Although 14,315 observations are removed from the data set, the binned residual plot does not change much (Figure 15). Therefore, I report results including these potential outliers (Table 5 for full results). Fixed effects of the hierarchical model are consistent with logistic model results. The model accuracy is 0.81, and the AUC is 0.79 (Figure 12). Due to page limitation, only variables of interests are reported and interpreted here (Table 3).

Table 3: Simple Hierarchical Model Results - Variables of Interests

Random effects:				
Groups	Name	Variance	Std.Dev.	
month	(Intercept)	0.1541	0.3925	
Number of obs: 41188, groups: month, 10				
Fixed effects:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1118102	0.1934678	-16.084	< 2e-16 ***
poutcomenonexistent	0.5002969	0.0865843	5.778	7.55e-09 ***
poutcomesuccess	1.7893375	0.0809860	22.094	< 2e-16 ***
confc	0.0842785	0.0076907	10.959	< 2e-16 ***
poutcomenonexistent:confc	-0.0290122	0.0079351	-3.656	0.000256 ***
poutcomesuccess:confc	-0.0342754	0.0114872	-2.984	0.002847 **

intercept: For a fixed month, the expected odds ratio of purchasing a term deposit is 0.04 ($e^{-3.112}$) if the consumer is below 30, divorced, working as an admin, with a basic 4-year education, and who has not been contacted right before this campaign and is approached on a Friday by cell phone for the first time for this campaign. This consumer has been approached for previous campaigns but has not bought any product. The daily Euribor 3 moth rate is high, but the consumer price index and consumer confidence index are on the average level.

poutcome: For a fixed month, keeping all other variables constant, compared to a consumer who did not agree to purchase a term deposit for previous campaigns, the odds ratio of purchasing a term deposit for this campaign increases by 5.98 ($e^{1.789}$) if the consumer has successfully purchased the product in previous campaigns, while the odds ratio only increases by 1.65 ($e^{0.500}$) if the consumer has not been approached in previous campaigns.

confc: For a fixed month, keeping all other variables constant, an additional unit increase in consumer confidence index increases the odds ratio of purchasing a term deposit for this campaign by 1.09 ($e^{0.084}$).

poutcome*confc: For a fixed month, keeping all other variables constant, when consumer confidence index increases by 1 unit, compared to a consumer who did not agree to purchase a term deposit in previous campaigns, the odds ratio of purchasing a term deposit for this campaign only increases by 1.05 ($e^{(0.084-0.034)}$) if the consumer has successfully purchased the product in previous campaign, and the odds ratio increases by 1.06 ($e^{(0.084-0.029)}$) if the consumer has never been approached before.

Random Effect - month: Taking April as an example: for any consumer contacted in April, the baseline odds of purchasing is 0.04 ($e^{(-3.112-0.167)}$), which is lower than the overall month wide average. The across-county variation attributed to the random intercept is 0.3925.

To answer the main research question, the odds ratio of consumers purchasing a term deposit for the current marketing campaign increases if they have successful experience in previous campaigns. However, the effect of previous campaign outcome diminishes if people are optimistic about future economy. Additionally, consumers are more likely to purchase a term deposit in a marketing campaign if they are contacted by cell phone on Monday and with successful outcomes in previous campaigns.

5. Conclusion

This analysis provides insights for banks to conduct successful marketing campaigns, especially for attracting term deposits. When people are pessimistic about future economy, banks should focus on loyal consumers and convince them to purchase term deposits again. When people are confident about future economy, banks could approach the general public for term deposits. Additionally, banks need to remember that consumers are less likely to purchase if they are contacted multiple times for the same marketing campaign.

One limitation of this analysis is that far more consumers in the dataset chose not to purchase a term deposit at the end of the marketing campaign, which may leads to potential bias in the analysis. Additionally, this dataset is solely based on one Portuguese bank. Future studies could explore the same research question with data from other countries and analyze if the impact of telephone marketing campaigns would be different.

Reference

S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems (2014), doi:10.1016/j.dss.2014.03.001.

Appendix

Table 4: Model Selection Process

Model	Predictors
M1(Baseline)	newage, job, marital, education, housing, loan, contact, month, day, campaign, previous, poutcome, loweuri, confc, pricec
M2(Interactions)	M1 + poutcome $*$ confc
M3(Stepwise)	newage, job, marital, education, contact, month, day, campaign, previous, poutcome, loweuri, confc, pricec, poutcome $*$ confc
M4(Hierarchical)	(1|month), newage, job, marital, education, contact, day, campaign, previous, poutcome, loweuri, confc, pricec, poutcome $*$ confc

* $log(\frac{\pi_{ij}}{1-\pi_{ij}})$ as response variable

*M4 is the final model

Table 5: Hierarchical Model Results

Random effects:				
Groups	Name	Variance	Std.Dev.	
month	(Intercept)	0.1541	0.3925	
Number of obs: 41188, groups:	month, 10			
Fixed effects:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1118102	0.1934678	-16.084	< 2e-16 ***
poutcomenonexistent	0.5002969	0.0865843	5.778	7.55e-09 ***
poutcomesuccess	1.7893375	0.0809860	22.094	< 2e-16 ***
contacttelephone	-0.5109690	0.0600645	-8.507	< 2e-16 ***
newage1	-0.1256645	0.0493806	-2.545	0.010933 *
newage2	0.1726695	0.1088135	1.587	0.112549
confc	0.0842785	0.0076907	10.959	< 2e-16 ***
pricec	0.5763663	0.0448670	12.846	< 2e-16 ***
loweuri1	2.1634644	0.0623028	34.725	< 2e-16 ***
previous	0.0442995	0.0531416	0.834	0.404500
jobblue-collar	-0.1766379	0.0688682	-2.565	0.010321 *
jobentrepreneur	-0.0624599	0.1066762	-0.586	0.558206
jobhousemaid	-0.1028875	0.1278346	-0.805	0.420907
jobmanagement	-0.0619316	0.0745173	-0.831	0.405915
jobretired	0.1285898	0.1007873	1.276	0.202008
jobself-employed	-0.0767367	0.1011959	-0.758	0.448272
jobservices	-0.1452526	0.0746407	-1.946	0.051652 .
jobstudent	0.1940960	0.0982516	1.976	0.048211 *
jobtechnician	-0.0124446	0.0615256	-0.202	0.839708
jobunemployed	-0.0009718	0.1100673	-0.009	0.992956
campaign	-0.0507620	0.0092471	-5.490	4.03e-08 ***
daymon	-0.2280943	0.0575171	-3.966	7.32e-05 ***
daythu	0.0475612	0.0553337	0.860	0.390045
daytue	0.0472444	0.0569329	0.830	0.406638
daywed	0.1349873	0.0565416	2.387	0.016968 *
educationbasic.6y	0.1416537	0.1006498	1.407	0.159311
educationbasic.9y	0.0108800	0.0799808	0.136	0.891796
educationhigh.school	0.1004085	0.0778517	1.290	0.197141
educationilliterate	0.6194988	0.5382881	1.151	0.249786
educationprofessional.course	0.0708652	0.0862707	0.821	0.411402
educationuniversity.degree	0.1918115	0.0779836	2.460	0.013908 *
maritalmarried	0.0351581	0.0596401	0.590	0.555523
maritalsingle	0.0521449	0.0670333	0.778	0.436630
poutcomenonexistent:confc	-0.0290122	0.0079351	-3.656	0.000256 ***
poutcomesuccess:confc	-0.0342754	0.0114872	-2.984	0.002847 **

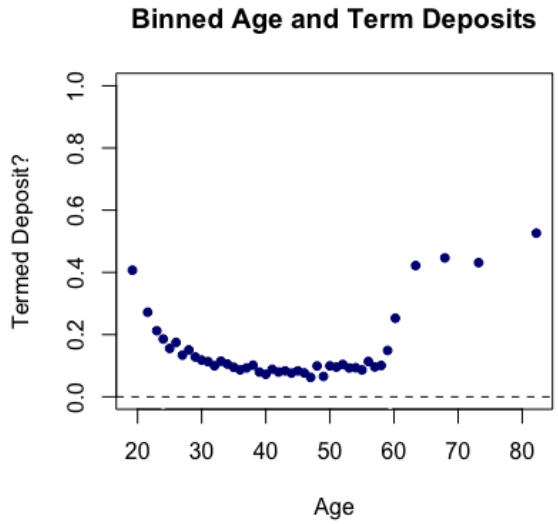

Figure 1: Binned Age and Term Deposits

Table 6: Job and Term Deposits

	admin.	blue-collar	entrepreneur	housemaid	management	retired
no	0.87	0.93	0.91	0.90	0.89	0.75
yes	0.13	0.07	0.09	0.10	0.11	0.25
	self-employed	services	student	technician	unemployed	unknown
no	0.90	0.92	0.69	0.89	0.86	
yes	0.10	0.08	0.31	0.11	0.14	

Table 7: Marriage and Term Deposits

	divorced	married	single	unknown
no	0.90	0.90	0.86	
yes	0.10	0.10	0.14	

Table 8: Education and Term Deposits

	basic.4y	basic.6y	basic.9y	high.school	illiterate
no	0.90	0.92	0.92	0.89	0.78
yes	0.10	0.08	0.08	0.11	0.22
	professional.course	university.degree	unknown		
no	0.89	0.86			
yes	0.11	0.14			

Table 9: Housing Loan and Term Deposits

	no	unknown	yes
no	0.89		0.88
yes	0.11		0.12

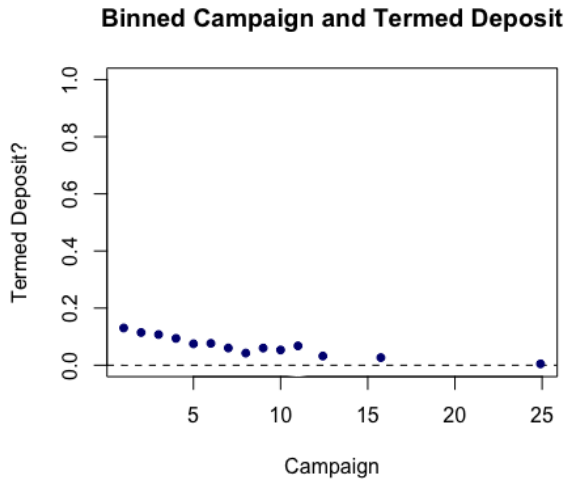

Figure 2: Binned Contacts for This Campaign and Term Deposits

Table 10: Personal Loan and Term Deposits

	no	unknown	yes
no	0.89		0.89
yes	0.11		0.11

Table 11: Contacted Type and Term Deposits

	cellular	telephone
no	0.85	0.95
yes	0.15	0.05

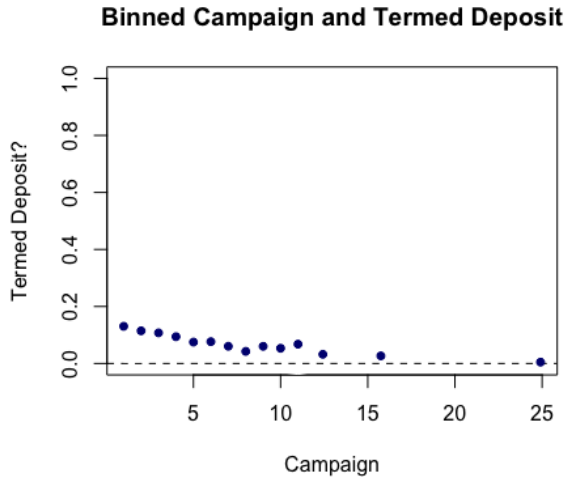

Figure 3: Binned Days since Last Contact and Term Deposits

Table 12: Contacted Month and Term Deposits

	apr	aug	dec	jul	jun	mar	may	nov	oct	sep
no	0.80	0.89	0.51	0.91	0.89	0.49	0.94	0.90	0.56	0.55
yes	0.20	0.11	0.49	0.09	0.11	0.51	0.06	0.10	0.44	0.45

Table 13: Contacted Day and Term Deposits

	fri	mon	thu	tue	wed
no	0.89	0.90	0.88	0.88	0.88
yes	0.11	0.10	0.12	0.12	0.12

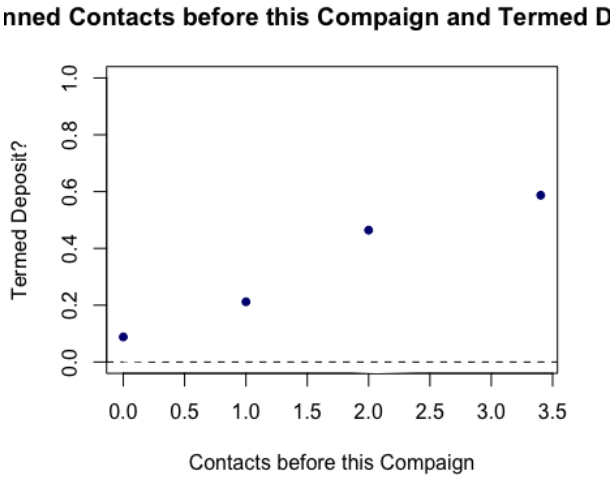

Figure 4: Binned Contacts before this Campaign and Term Deposits

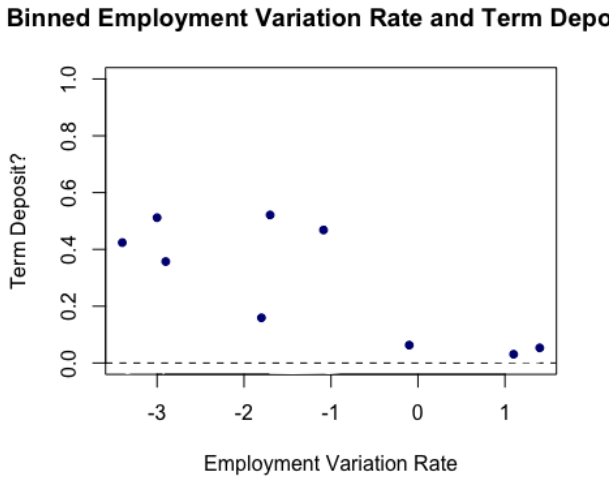

Figure 5: Binned Employment Variation Rate and Term Deposits

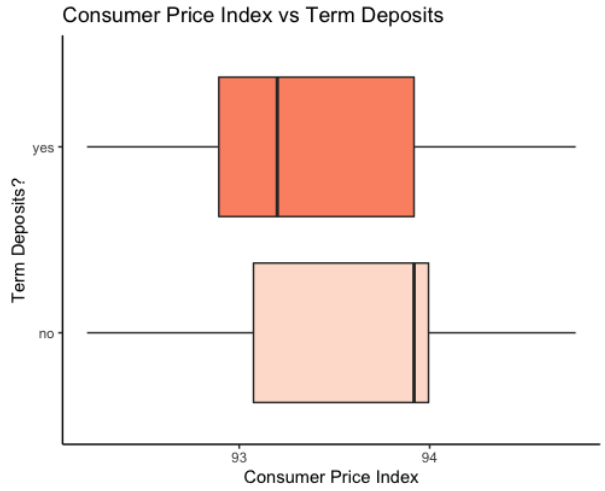

Figure 6: Binned Consumer Price Index and Term Deposits

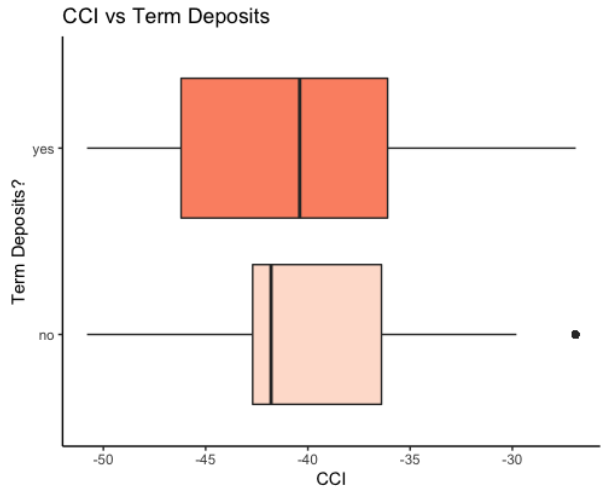

Figure 7: Binned Consumer Confidence Index and Term Deposits

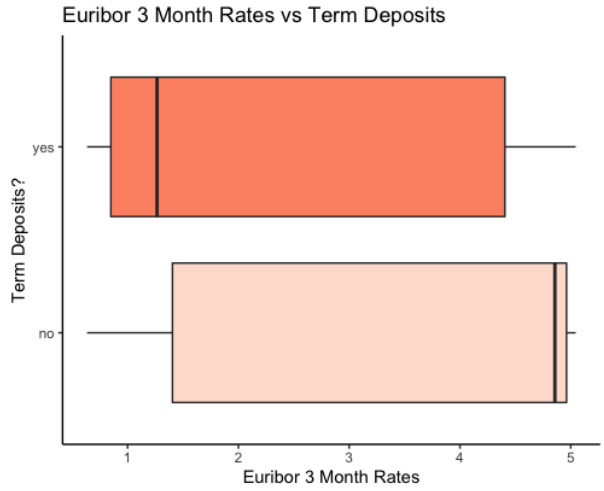

Figure 8: Binned Euribor 3 Month Rates and Term Deposits

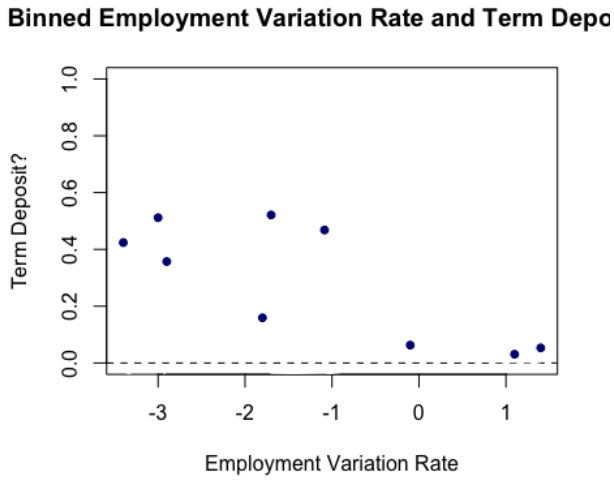

Figure 9: Binned Employment Variation Rate and Term Deposits

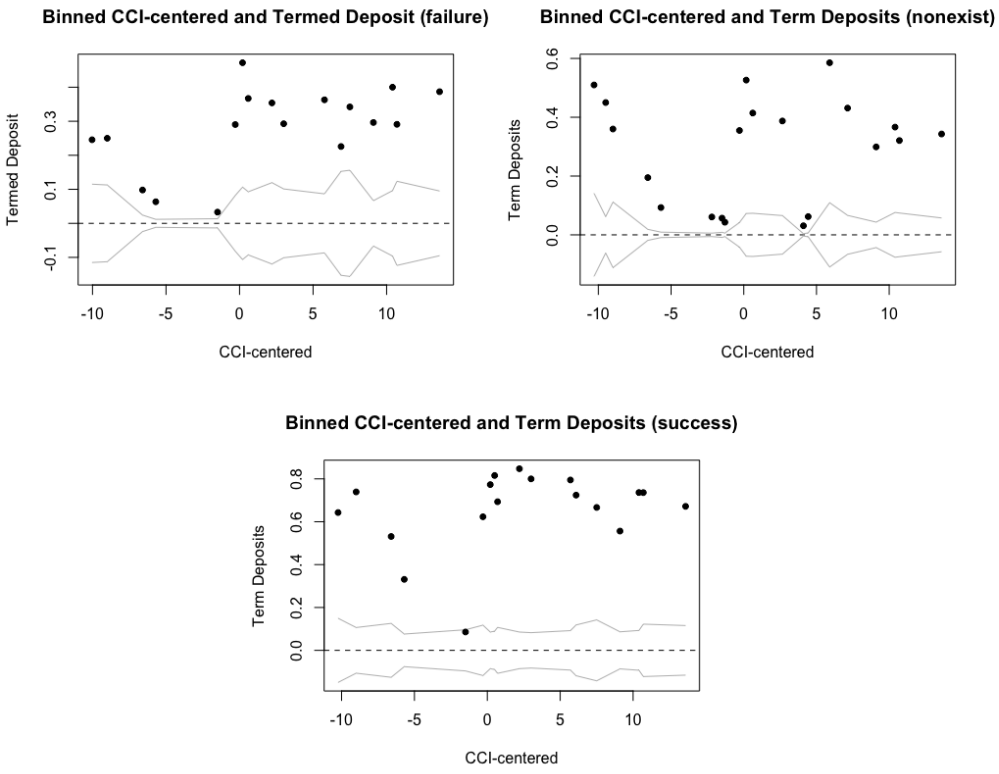

Figure 10: Interaction: CCI and Previous Campaign Outcome

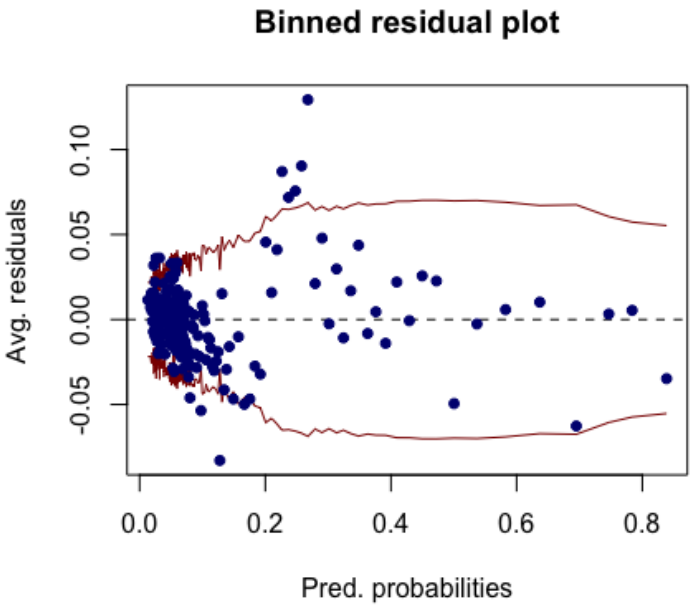

Figure 11: Binned Residual Plot

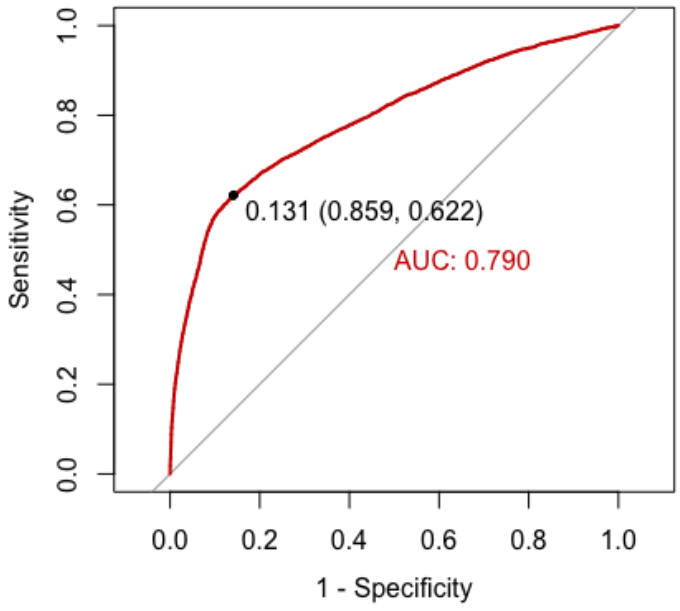

Figure 12: AUC

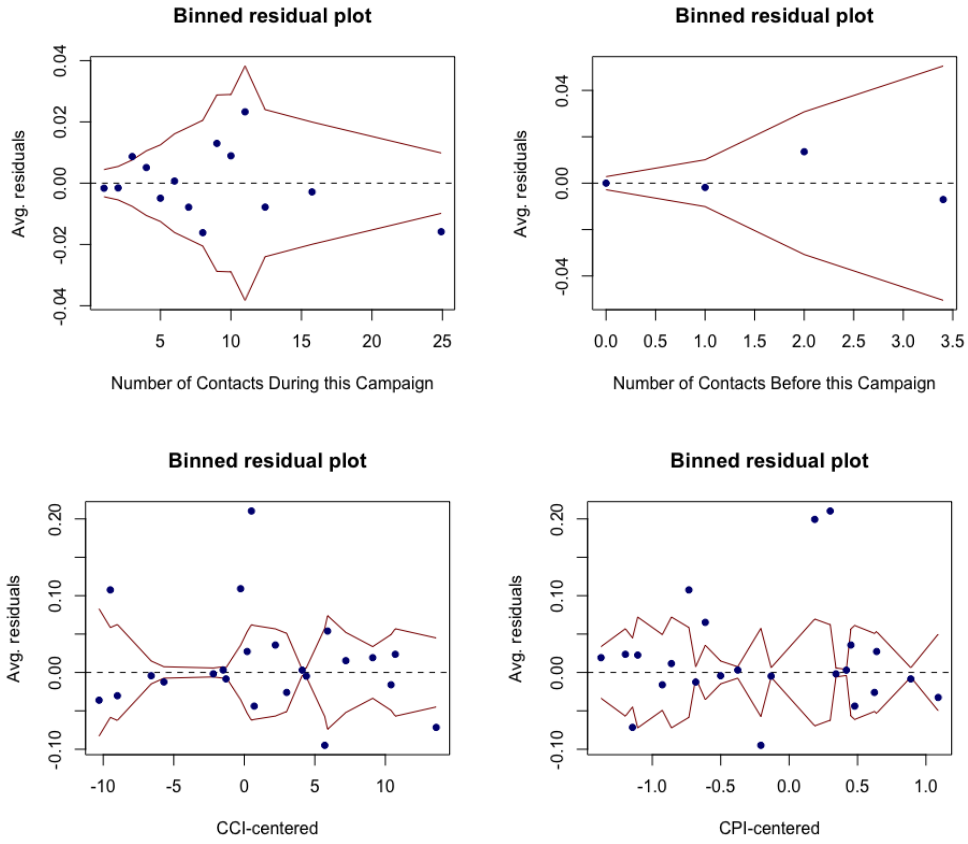

Figure 13: Linearity Check

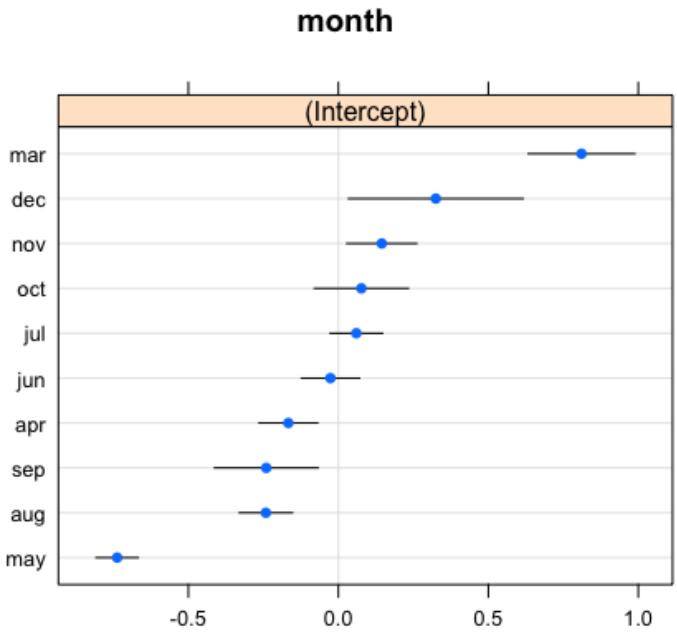

Figure 14: Outlier Check

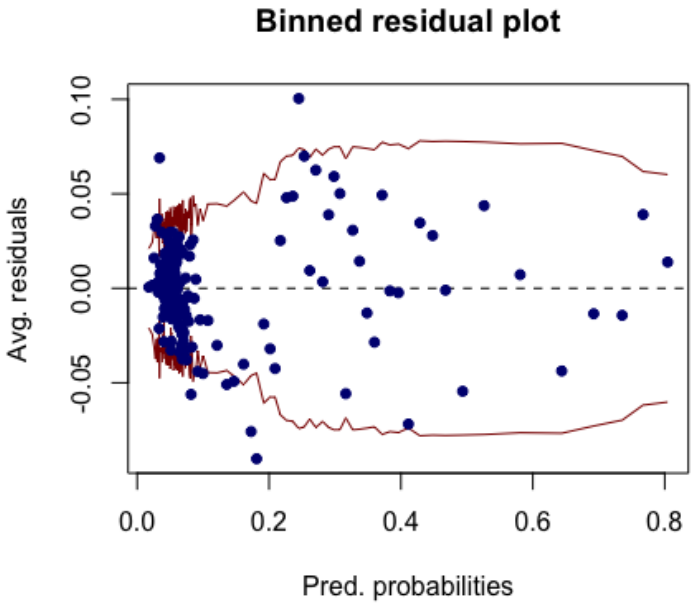

Figure 15: Binned Residual Plot without Potential Outliers

Main Code

Data Prepare

```
# Data Loading
bank <- read.csv(/Users/Chenyu/Downloads/bank-additional-full.csv, header=TRUE, stringsAsFactors=FALSE)

bank$job <- as.factor(bank$job)
bank$marital <- as.factor(bank$marital)
bank$education <- as.factor(bank$education)
bank$housing <- as.factor(bank$housing)
bank$loan <- as.factor(bank$loan)
bank$contact <- as.factor(bank$contact)
bank$month <- as.factor(bank$month)
bank$day_of_week <- as.factor(bank$day_of_week)
bank$poutcome <- as.factor(bank$poutcome)
bank$y <- as.factor(bank$y)
bank$ynum <- as.numeric(bank$y) - 1

# install.packages(naniar)
library(naniar)
bankna <- replace_with_na(bank, replace = list(job =unknown, marital = unknown, education = unknown,
                                               housing = unknown, loan = unknown, poutcome = unknown))
na_count <-sapply(bankna, function(y) sum(length(which(is.na(y)))))
na_count <- data.frame(na_count)
bank = subset(bankna, select = -c(default,duration))
summary(bank)

library(mice)
# Visualize Missing Data
md.pattern(bank)
# install.packages(VIM)
library(VIM);library(lattice)
aggr(bank,col=c(lightblue3,darkred),numbers=TRUE,sortVars=TRUE,labels=names(bank), cex.axis=.7,gap=3,ylab=c(Proportion missing,Missingness pattern))
marginplot(bank[,c(diameter.,age)],col=c(lightblue3,darkred),cex.numbers = 1.2,pch =19)
# Imputation
bank_imp <- mice(bank,m=5,print=F)

library(ggplot2)
```

```
d1 <- complete(bank_imp,1);d1
apply(table(d1[,c(y,education)])/sum(table(d1[,c(y,education)])),2,function(x) x/
    sum(x))
apply(table(d1[,c(y,housing)])/sum(table(d1[,c(y,housing)])),2,function(x) x/sum(x
    ))
apply(table(d1[,c(y,loan)])/sum(table(d1[,c(y,loan)])),2,function(x) x/sum(x))
d1$oldconsumer[d1$pdays == 999] <- 0
d1$oldconsumer[d1$pdays < 999] <- 1
table(d1$oldconsumer)
d1$newage[d1$age <=30] <- 0
d1$newage[d1$age >30 & d1$age < 60] <- 1
d1$newage[d1$age >=60] <- 2
d1$newage <- as.factor(d1$newage)
table(d1$newage)

d2 <- complete(bank_imp,2);d2
apply(table(d2[,c(y,education)])/sum(table(d2[,c(y,education)])),2,function(x) x/
    sum(x))
apply(table(d2[,c(y,housing)])/sum(table(d2[,c(y,housing)])),2,function(x) x/sum(x
    ))
apply(table(d2[,c(y,loan)])/sum(table(d2[,c(y,loan)])),2,function(x) x/sum(x))

bank_obs <- na.omit(bank)
apply(table(bank_obs[,c(y,education)])/sum(table(bank_obs[,c(y,education)])),2,
    function(x) x/sum(x))
apply(table(bank_obs[,c(y,housing)])/sum(table(bank_obs[,c(y,housing)])),2,
    function(x) x/sum(x))
apply(table(bank_obs[,c(y,loan)])/sum(table(bank_obs[,c(y,loan)])),2,function(x) x
    /sum(x))

library(MatchIt) #for propensity score matching
library(cobalt)
library(tidyverse)
library(knitr)
library(GGally)
library(xtable)
library(arm)
library(pROC)
library(e1071)
library(caret)
library(rms)
require(gridExtra)

# EDA
```

```
73
74 # age (none / people are less likely to purchase the product during their mid-ages
       (30s-50s), compared to their 20s and 60s)
75 table(bank$age)
76 ggplot(bank,aes(x=y, y=age, fill=y)) +
77   geom_boxplot() + coord_flip() +
78   scale_fill_brewer(palette=Reds) +
79   labs(title=Age vs Termed Deposit,
80        x=Termed Deposit?,y=Age) +
81   theme_classic() + theme(legend.position=none)
82 binnedplot(y=bank$ynum,bank$age,xlab=Age,ylim=c(0,1),col.pts=navy,
83            ylab =Termed Deposit?,main=Binned Age and Term Deposits,
84            col.int=white)
85
86
87 # job (retired - highest)
88 table(bank$job)
89 tjob <- apply(table(bank[,c(y,job)])/sum(table(bank[,c(y,job)])),2,function(x) x/
       sum(x))
90 xtable(tjob)
91 plot(0:11,tapply(bank$ynum, bank$job, mean),col='blue4',pch=10)
92
93 # marital (similar)
94 table(bank$marital)
95 tmarital <- apply(table(bank[,c(y,marital)])/sum(table(bank[,c(y,marital)])),2,
       function(x) x/sum(x))
96 xtable(tmarital)
97 plot(0:3,tapply(bank$ynum, bank$marital, mean),col='blue4',pch=10)
98
99 # education (high school - highest)
100 table(bank$education)
101 teducation <- apply(table(bank[,c(y,education)])/sum(table(bank[,c(y,education)]))
       ,2,function(x) x/sum(x))
102 xtable(teducation)
103 plot(0:7,tapply(bank$ynum, bank$education, mean),col='blue4',pch=10)
104
105 # housing (similar)
106 table(bank$housing)
107 thousing <- apply(table(bank[,c(y,housing)])/sum(table(bank[,c(y,housing)])),2,
       function(x) x/sum(x))
108 xtable(thousing)
109 plot(0:2,tapply(bank$ynum, bank$housing, mean),col='blue4',pch=10)
110
111 # loan (similar)
```

```
table(bank$loan)
tloan <- apply(table(bank[,c(y,loan)])/sum(table(bank[,c(y,loan)])),2,function(x)
    x/sum(x))
xtable(tloan)
plot(0:2,tapply(bank$ynum, bank$loan, mean),col='blue4',pch=10)

# contact (cellular slightly higher)
tcontact <- apply(table(bank[,c(y,contact)])/sum(table(bank[,c(y,contact)])),2,
    function(x) x/sum(x))
xtable(tcontact)
plot(0:1,tapply(bank$ynum, bank$contact, mean),col='blue4',pch=10)

# month (apr, aug, jul, jun, may, nov)
tmonth <- apply(table(bank[,c(y,month)])/sum(table(bank[,c(y,month)])),2,function(
    x) x/sum(x))
xtable(tmonth)
plot(0:9,tapply(bank$ynum, bank$month, mean),col='blue4',pch=10)

# day (similar)
plot(0:4,tapply(bank$ynum, bank$day_of_week, mean),col='blue4',pch=10)

# duration (positive) ? don't include
ggplot(bank,aes(x=y, y=duration, fill=y)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette=Reds) +
  labs(title=Duration vs Termed Deposit,
       x=Termed Deposit?,y=Duration) +
  theme_classic() + theme(legend.position=none)
binnedplot(y=bank$ynum,bank$duration,xlab=Duration,ylim=c(0,1),col.pts=navy,
           ylab =Termed Deposit?,main=Binned Duration and Termed Deposit,
           col.int=white)
cor(bank$ynum,bank$duration)
regduration <- glm(y~duration, data=bank, family = binomial)
summary(regduration)

# campaign (none / decreasing over the number of contacts)
ggplot(bank,aes(x=y, y=campaign, fill=y)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette=Reds) +
  labs(title=Campaign vs Termed Deposit,
       x=Termed Deposit?,y=Campaign) +
  theme_classic() + theme(legend.position=none)
binnedplot(y=bank$ynum,bank$campaign,xlab=Campaign,ylim=c(0,1),col.pts=navy,
           ylab =Term Deposits?,main=Binned Number of Contacts During this
               Campaign and Term Deposits,
```

```
153           col.int=white)
154
155 # previous
156 ggplot(bank,aes(x=y, y=previous, fill=y)) +
157   geom_boxplot() + coord_flip() +
158   scale_fill_brewer(palette=Reds) +
159   labs(title=Contacts for this Compaign vs Termed Deposit,
160        x=Termed Deposit?,y=Contacts for this Compaign) +
161   theme_classic() + theme(legend.position=none)
162 bank$yescontact[bank$previous == 0] <- 0
163 bank$yescontact[bank$previous != 0] <- 1
164 tyescontact <- apply(table(bank[,c(y,yescontact)])/sum(table(bank[,c(y,yescontact)
    ])),2,function(x) x/sum(x))
165 xtable(tyescontact)
166 binnedplot(y=bank$ynum,bank$previous,xlab=Contacts before this
    Campaign,ylim=c(0,1),col.pts=navy,
167           ylab =Term Deposit?,main=Binned Contacts before this Campaign and Term
                Deposits,
168           col.int=white)
169 # plot(0:1,tapply(bank$ynum, bank$yescontact, mean),col='blue4',pch=10)
170
171 # poutcome: consumers who have previously bought termed deposits are more likely
    to purchase again
172 table(bank$poutcome)
173 tpoutcome <- apply(table(bank[,c(y,poutcome)])/sum(table(bank[,c(y,poutcome)])),2,
    function(x) x/sum(x))
174 xtable(tpoutcome)
175 plot(0:2,tapply(bank$ynum, bank$poutcome, mean),col='blue4',pch=10)
176
177 # employment variation rate (negative) [drop - high correlation with CPI and
    Euribor]
178 ggplot(bank,aes(x=y, y=emp.var.rate, fill=y)) +
179   geom_boxplot() + coord_flip() +
180   scale_fill_brewer(palette=Reds) +
181   labs(title=Employment Variation Rate vs Termed Deposit,
182        x=Termed Deposit?,y=Employment Variation Rate) +
183   theme_classic() + theme(legend.position=none)
184 binnedplot(y=bank$ynum,bank$emp.var.rate,xlab=Employment Variation
    Rate,ylim=c(0,1),col.pts=navy,
185           ylab =Term Deposit?,main=Binned Employment Variation Rate and Term
                Deposits,
186           col.int=white)
187
188 # consumer price index (negative)
189 ggplot(bank,aes(x=y, y=cons.price.idx, fill=y)) +
190   geom_boxplot() + coord_flip() +
```

```
  scale_fill_brewer(palette=Reds) +
  labs(title=Consumer Price Index vs Term Deposits,
       x=Term Deposits?,y=Consumer Price Index) +
  theme_classic() + theme(legend.position=none)
binnedplot(y=bank$ynum,bank$cons.price.idx,xlab=Consumer Price Index,ylim=c(0,1),col.pts=navy,
           ylab =Term Deposits?,main=Binned Consumer Price Index and Term Deposits,
           col.int=white)

# CCI (positive)
ggplot(bank,aes(x=y, y=cons.conf.idx, fill=y)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette=Reds) +
  labs(title=CCI vs Term Deposits,
       x=Term Deposits?,y=CCI) +
  theme_classic() + theme(legend.position=none)
binnedplot(y=bank$ynum,bank$cons.conf.idx,xlab=CCI,ylim=c(0,1),col.pts=navy,
           ylab =Term Deposits?,main=Binned CCI and Term Deposits,
           col.int=white)

# Euribor (negative)
ggplot(bank,aes(x=y, y=euribor3m, fill=y)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette=Reds) +
  labs(title=Euribor 3 Month Rates vs Term Deposits,
       x=Term Deposits?,y=Euribor 3 Month Rates) +
  theme_classic() + theme(legend.position=none)
binnedplot(y=bank$ynum,bank$euribor3m,xlab=Euribor 3 Month Rates,ylim=c(0,1),col.pts=navy,
           ylab =Term Deposits?,main=Binned Euribor 3 Month Rates and Term Deposits,
           col.int=white)
summary(d1$euribor3m)
d1$loweuri[d1$euribor3m<=3]<-1
d1$loweuri[d1$euribor3m>3]<-0
d1$loweuri<-as.factor(d1$loweuri)

# correlation: employment variation rate, consumer price index, CCI, Euribor
cor(bank$emp.var.rate, bank$cons.conf.idx)
cor(bank$cons.price.idx, bank$emp.var.rate)
cor(bank$emp.var.rate, bank$euribor3m)
cor(bank$cons.price.idx, bank$cons.conf.idx)
cor(bank$cons.price.idx, bank$euribor3m)
cor(bank$cons.conf.idx, bank$euribor3m)

```

```
# interaction
# age * poutcome
binnedplot(d1$agec[d1$poutcome==failure],
           y=d1$ynum[d1$poutcome==failure],
           xlab = Age-centered, ylab = Termed Deposit, main = Binned Age-centered
               and Termed Deposit (failure))
binnedplot(d1$agec[d1$poutcome==nonexistent],
           y=d1$ynum[d1$poutcome==nonexistent],
           xlab = Age-centered, ylab = Termed Deposit, main = Binned Age-centered
               and Termed Deposit (nonexist))
binnedplot(d1$agec[d1$poutcome==success],
           y=d1$ynum[d1$poutcome==success],
           xlab = Age-centered, ylab = Termed Deposit, main = Binned Age-centered
               and Termed Deposit (success))

# newage:poutcome
plot(0:2,tapply(d1$ynum[d1$poutcome==failure], d1$newage[d1$poutcome==failure],
    mean),col='blue4',pch=10,ann=FALSE,ylim=c(0,1),cex=1.5)
par(new=TRUE)
plot(0:2,tapply(d1$ynum[d1$poutcome==nonexistent], d1$newage[d1$poutcome==
    nonexistent], mean),col='red4',pch=10,ann=FALSE,ylim=c(0,1),cex=1.5)
par(new=TRUE)
plot(0:2,tapply(d1$ynum[d1$poutcome==success], d1$newage[d1$poutcome==success],
    mean),col='black',pch=10,ann=FALSE,ylim=c(0,1),cex=1.5)
title(xlab = Age,ylab = 'Prob of Purchasing a Term Deposit')
legend(topright,pch=c(10,10),legend=c(failure,nonexistent,success),col=c(blue4,
    red4,black),bty=n)

# poutcome:confc
binnedplot(d1$confc[d1$poutcome==failure],
           y=d1$ynum[d1$poutcome==failure],
           xlab = CCI-centered, ylab = Term Deposits, main = Binned CCI-centered
               and Term Deposits (failure))
binnedplot(d1$confc[d1$poutcome==nonexistent],
           y=d1$ynum[d1$poutcome==nonexistent],
           xlab = CCI-centered, ylab = Term Deposits, main = Binned CCI-centered
               and Term Deposits (nonexist))
binnedplot(d1$confc[d1$poutcome==success],
           y=d1$ynum[d1$poutcome==success],
           xlab = CCI-centered, ylab = Term Deposits, main = Binned CCI-centered
               and Term Deposits (success))

# poutcome:loweuri
plot(0:1,tapply(d1$ynum[d1$poutcome==failure], d1$loweuri[d1$poutcome==failure],
    mean),col='blue4',pch=10,ann=FALSE,ylim=c(0,1),cex=1.5)
par(new=TRUE)
```

```
plot(0:1,tapply(d1$ynum[d1$poutcome==nonexistent], d1$loweuri[d1$poutcome==
    nonexistent], mean),col='red4',pch=10,ann=FALSE,ylim=c(0,1),cex=1.5)
par(new=TRUE)
plot(0:1,tapply(d1$ynum[d1$poutcome==success], d1$loweuri[d1$poutcome==success],
    mean),col='black',pch=10,ann=FALSE,ylim=c(0,1),cex=1.5)
title(xlab = Euribor,ylab = 'Prob of Purchasing a Term Deposit')
legend(topright,pch=c(10,10),legend=c(failure,nonexistent,success),col=c(blue4,
    red4,black),bty=n)
```

Model

```
# centering
bank$agec <- bank$age-mean(bank$age)
bank$varc <- bank$emp.var.rate + mean(bank$emp.var.rate)
bank$pricec <- bank$cons.price.idx - mean(bank$cons.price.idx)
bank$confc <- bank$cons.conf.idx - mean(bank$cons.conf.idx)
bank$euric <- bank$euribor3m - mean(bank$euribor3m)

d1$agec <- d1$age-mean(d1$age)
d1$varc <- d1$emp.var.rate + mean(d1$emp.var.rate)
d1$pricec <- d1$cons.price.idx - mean(d1$cons.price.idx)
d1$confc <- d1$cons.conf.idx - mean(d1$cons.conf.idx)d
d1$euric <- d1$euribor3m - mean(d1$euribor3m)

# model 1
model1 <- glm(ynum~ newage + job + marital + education + housing + loan + contact
              + month + day_of_week + campaign + previous + poutcome
              + varc + pricec + confc + loweuri, data = d1, family = binomial)
summary(model1)

rawresid1 <- residuals(model1,resp)
binnedplot(x=fitted(model1),y=rawresid1,xlab=Pred.  probabilities,
           col.int=red4,ylab=Avg.  residuals,main=Binned residual
               plot,col.pts=navy)
binnedplot(x=d1$campaign,y=rawresid1,xlab=Number of Contacts During this Campaign,
           col.int=red4,ylab=Avg.  residuals,main=Binned residual
               plot,col.pts=navy)
binnedplot(x=d1$previous,y=rawresid1,xlab=Contacts for this Campaign,
           col.int=red4,ylab=Avg.  residuals,main=Binned residual
               plot,col.pts=navy)
binnedplot(x=d1$varc,y=rawresid1,xlab=employment variation rate centered,
           col.int=red4,ylab=Avg.  residuals,main=Binned residual
               plot,col.pts=navy)
binnedplot(x=d1$pricec,y=rawresid1,xlab=CPI centered,
```

```
30            col.int=red4,ylab=Avg.  residuals,main=Binned residual
                  plot,col.pts=navy)
31  binnedplot(x=d1$confc,y=rawresid1,xlab=CCI centered,
32            col.int=red4,ylab=Avg.  residuals,main=Binned residual
                  plot,col.pts=navy)
33
34  # Model validation
35  # multicollinearity
36  vif(model1)
37  #let's do the confusion matrix with .5 threshold
38  Conf_mat <- confusionMatrix(as.factor(ifelse(fitted(model1) >=  mean(d1$ynum), 1,
        0)),
39                              as.factor(d1$ynum),positive = 1)
40  Conf_mat$table
41  Conf_mat$overall[Accuracy];
42  Conf_mat$byClass[c(Sensitivity,Specificity)] #True positive rate and True negative
        rate
43  #Maybe we can try to increase that accuracy.
44  roc(bank$ynum,fitted(model1),plot=T,print.thres=best,legacy.axes=T,
45      print.auc =T,col=red3)
46
47  model2 <- glm(ynum~ newage + job + marital + education + housing + loan + previous
48                + contact + month + day_of_week + campaign + poutcome + confc +
                      loweuri,
49                data = d1, family = binomial)
50  summary(model2)
51  vif(model2)
52
53  # interaction
54  model3 <- glm(ynum~ newage + job + marital + education + housing + loan + previous
55                + contact + month + day_of_week + campaign + poutcome
56                + confc + pricec + loweuri + poutcome:confc, data = d1, family =
                      binomial)
57  summary(model3)
58  vif(model3)
59
60  # Stepwise
61  model0 <- glm(ynum~ 1, data = d1, family = binomial)
62  model_stepwiseaic <- step(model0,scope=formula(model3),direction=both, trace=0)
63  model_stepwiseaic$call
64  Model_aic <- glm(ynum~ loweuri + month + poutcome + confc + pricec +
65                     contact + day_of_week + job + campaign + newage + education +
66                     poutcome:confc,
67                   data = d1, family = binomial)
68  summary(Model_aic)
```

```
anova(model3, Model_aic, test= Chisq)
# no difference: pick Model_aic

vif(Model_aic)
rawresid2 <- residuals(Model_aic,resp)
binnedplot(x=fitted(Model_aic),y=rawresid2,xlab=Pred.  probabilities,
           col.int=red4,ylab=Avg.  residuals,main=Binned residual
               plot,col.pts=navy)
binnedplot(x=d1$previous,y=rawresid2,xlab=Contacts for this Campaign,
           col.int=red4,ylab=Avg.  residuals,main=Binned residual
               plot,col.pts=navy)
binnedplot(x=d1$confc,y=rawresid2,xlab=CCI centered,
           col.int=red4,ylab=Avg.  residuals,main=Binned residual
               plot,col.pts=navy)
binnedplot(x=d1$pricec,y=rawresid2,xlab=CPI centered,
           col.int=red4,ylab=Avg.  residuals,main=Binned residual
               plot,col.pts=navy)
binnedplot(x=d1$campaign,y=rawresid2,xlab=Number of Contacts During this Campaign,
           col.int=red4,ylab=Avg.  residuals,main=Binned residual
               plot,col.pts=navy)

Conf_mat <- confusionMatrix(as.factor(ifelse(fitted(Model_aic) >=  mean(d1$ynum),
    1,0)),
                            as.factor(d1$ynum),positive = 1)
Conf_mat$table
Conf_mat$overall[Accuracy];
Conf_mat$byClass[c(Sensitivity,Specificity)] #True positive rate and True negative
     rate
#Maybe we can try to increase that accuracy.
roc(d1$ynum,fitted(Model_aic),plot=T,print.thres=best,legacy.axes=T,
    print.auc =T,col=red3)

# CPI
model4 <- glm(ynum~ poutcome + month + contact + newage + pricec +
                job + campaign + day_of_week + education + marital+poutcome:pricec
                    ,
              data = d1, family = binomial)
vif(model4)
rawresid4 <- residuals(model4,resp)
binnedplot(x=fitted(model4),y=rawresid4,xlab=Pred.  probabilities,
           col.int=red4,ylab=Avg.  residuals,main=Binned residual
               plot,col.pts=navy)
binnedplot(x=d1$pricec,y=rawresid4,xlab=CPI centered,
           col.int=red4,ylab=Avg.  residuals,main=Binned residual
               plot,col.pts=navy)

```

```
model5 <- glm(ynum~ poutcome + month + contact + newage + confc + pricec + loweuri
    + previous +
                job + campaign + day_of_week + education + marital+ poutcome:confc
                    ,
              data = d1, family = binomial)
summary(model5)
vif(model5)
rawresid5 <- residuals(model5,resp)
binnedplot(x=fitted(model5),y=rawresid5,xlab=Pred.  probabilities,
           col.int=red4,ylab=Avg.  residuals,main=Binned residual
               plot,col.pts=navy)
binnedplot(x=d1$confc,y=rawresid5,xlab=CCI centered,
           col.int=red4,ylab=Avg.  residuals,main=Binned residual
               plot,col.pts=navy)
binnedplot(x=d1$pricec,y=rawresid5,xlab=CPI centered,
           col.int=red4,ylab=Avg.  residuals,main=Binned residual
               plot,col.pts=navy)
binnedplot(x=d1$campaign,y=rawresid5,xlab=Number of Contacts During this Campaign,
           col.int=red4,ylab=Avg.  residuals,main=Binned residual
               plot,col.pts=navy)


# Hierachical Model
multimodel5 <- glmer(formula = ynum~poutcome + (1|month) + contact + newage +
    confc + pricec + loweuri + previous +
                       job + campaign + day_of_week + education + marital +
                           poutcome:confc, family = binomial(link=logit),
                     data = d1)
summary(multimodel5)
library(sjPlot)
tab_model(multimodel5)
dotplot(ranef(multimodel5, condVar=TRUE))

rawresid5 <- residuals(multimodel5,resp)
binnedplot(x=fitted(multimodel5),y=rawresid5,xlab=Pred.  probabilities,
           col.int=red4,ylab=Avg.  residuals,main=Binned residual
               plot,col.pts=navy)
binnedplot(x=d1$previous,y=rawresid5,xlab=Number of Contacts Before this Campaign,
           col.int=red4,ylab=Avg.  residuals,main=Binned residual
               plot,col.pts=navy)
binnedplot(x=d1$confc,y=rawresid5,xlab=CCI-centered,
           col.int=red4,ylab=Avg.  residuals,main=Binned residual
               plot,col.pts=navy)
binnedplot(x=d1$pricec,y=rawresid5,xlab=CPI-centered,
           col.int=red4,ylab=Avg.  residuals,main=Binned residual
               plot,col.pts=navy)
binnedplot(x=d1$campaign,y=rawresid5,xlab=Number of Contacts During this Campaign,
```

```
          col.int=red4,ylab=Avg.  residuals,main=Binned residual
              plot,col.pts=navy)

(ranef(multimodel5)$month)[apr,]

Conf_mat <- confusionMatrix(as.factor(ifelse(fitted(multimodel5) >=  mean(d1$ynum)
    , 1,0)),
                            as.factor(d1$ynum),positive = 1)
Conf_mat$table
Conf_mat$overall[Accuracy];
Conf_mat$byClass[c(Sensitivity,Specificity)] #True positive rate and True negative
     rate
#Maybe we can try to increase that accuracy.
roc(d1$ynum,fitted(multimodel5),plot=T,print.thres=best,legacy.axes=T,
    print.auc =T,col=red3)

# Eliminate Outlier: May & March
newd1 <- d1[!(d1$month==may),]
newd1 <- newd1[!(newd1$month==mar),]
newmodel <- glmer(formula = ynum~poutcome + (1|month) + contact + newage + confc +
     pricec + loweuri + previous +
                   job + campaign + day_of_week + education + marital + poutcome:
                       confc, family = binomial(link=logit),
                  data = newd1)
summary(newmodel)
newrawresid <- residuals(newmodel,resp)
binnedplot(x=fitted(newmodel),y=newrawresid,xlab=Pred.  probabilities,
          col.int=red4,ylab=Avg.  residuals,main=Binned residual
              plot,col.pts=navy)
```