Syndicate 5 Statistical Learning Problem Set #1

May 31, 2021

Credit Card Balances

Question 1

The expected relationship roughly shows the common sense and logical judgement about the potential relationship between the credit card balance and independent parameters.

It is assumed that the credit limit, rating, cards on hand, user's age and the identity of student could be the main factors that affect the balance of the credit card, which is shown in the following table. Based on these assumptions, detailed regression analysis will be conducted and explained in the following paragraphs.

Question 2

A multiple linear regression was run with all possible predictors holding 'balance' as the outcome and other variables as predictors. Predictors 'gender', 'student', 'married', and 'ethnicity' were assumed to be categorical with associated dummy variables created for each predictor.

Using R, the following coefficients in Table 1 were estimated.

Table 1: Multiple Linear Regression

	Dependent variable:
	Balance
Income	-7.718^{***}
	(0.241)
Limit	0.189^{***}
	(0.033)
Rating	1.150^{**}
	(0.498)
Cards	17.720^{***}
	(4.413)
Age	-0.558^{*}
	(0.302)

Education6	-51.878
	(109.201)
Education7	-45.627
	(105.953)
Education8	-66.490
	(103.493)
Education9	-7.671
	(101.730)
Education10	-21.790
	(101.753)
Education11	-63.956
	(101.285)
Education12	-42.985
	(100.883)
Education13	-43.381
	(101.073)
Education14	-23.939
	(100.841)
Education15	-36.608
	(100.715)
Education16	-30.359
	(100.698)
Education17	-44.087
	(101.368)
Education18	-79.711
	(101.868)
Education19	-44.960
	(104.458)
Education20	-97.893
	(122.352)
GenderFemale	-7.589
	(10.058)

StudentYes	422.261***
	(17.057)
MarriedYes	−8.585
	(10.528)
EthnicityAsian	16.147
	(14.463)
EthnicityCaucasian	9.846
	(12.445)
Constant	−457.639***
	(104.087)
Observations	400
R^2	0.957
Adjusted R^2	0.954
Residual Std. Error	99.037 (df = 374)
F Statistic	328.993*** (df = 25; 374)
Note:	*p<0.1; **p<0.05; ***p<0.01

Question 3

Our first model is the same as Q2 except with the *Education* variable now grouped into three bins based on the likely degree attained. This is to make the regression easier to interpret. The resultant regression is quite good with an Adjusted R^2 of 0.954 but we still have many insignificant predictors. However, before removing these, we want to use the second model to avoid any of the common pitfalls of regression.

Looking at a correlation matrix of numeric variables, we can see that *Rating* and *Limit* are extremely highly correlated and may give us multicollinearity issues. We conducted a Wald test to see if these predictors are jointly useful in predicting *Balance*.

$$H_0 : \beta_{Rating} = \beta_{Limit}$$

$$H_1 : \beta_{Rating} \neq \beta_{Limit}$$

The Wald test gives us a p-value of 0.062 which is not enough evidence to suggest that coefficients for *Rating* and *Limit* are jointly useful (at the 5% level). We decided to keep *Rating* and drop *Limit* for our second model since it is more likely to dictate the value for each *Limit* than vice versa, and thus more useful.

Before constructing our second model, we first looked at the residuals from Q2 and we see some non-linear patterns in *Income* and *Rating* so we chose to include squared terms for our second model. We considered the possibility of interaction terms for this regression however, we couldn't see any likely case the effect of a statistically significant independent variable on the dependent variable would depend on another (statistically significant) independent variable. This model is again quite good with an Adjusted R^2 of 0.963, and our new terms being statistically significant (interestingly, with the omission of *Limit*, *Cards* is no longer statistically significant), but we still have many insignificant coefficients which we can drop to simplify the regression at little cost to accuracy.

Thus our third model only has significant terms (all significant at the 1% level) and maintains the same very high Adjusted R^2 of 0.963 with the Residual Std. Error only increasing slightly (88.081 to 88.218). While both model 2 and 3 have similar accuracy, model 3 is preferred due to having fewer terms and thus being easier to interpret.

$$\text{Final Model: } Balance_i = -329.576 - 6.238 Income_i - 0.021 Income_i^2 + 2.471 Rating_i + 0.002 Rating_i^2 - 0.0729 Age_i + 428.341 Student_i$$

Table 2: Correlation Matrix (numeric data only)

	Income	Limit	Rating	Cards	Age	Balance
Income	1	0.792	0.791	-0.018	0.175	0.464
Limit	0.792	1	0.997	0.010	0.101	0.862
Rating	0.791	0.997	1	0.053	0.103	0.864
Cards	-0.018	0.010	0.053	1	0.043	0.086
Age	0.175	0.101	0.103	0.043	1	0.002
Balance	0.464	0.862	0.864	0.086	0.002	1

Table 3: Wald Test for Rating and Limit

Wald stat	p-value
3.487	0.062

Residual Plots

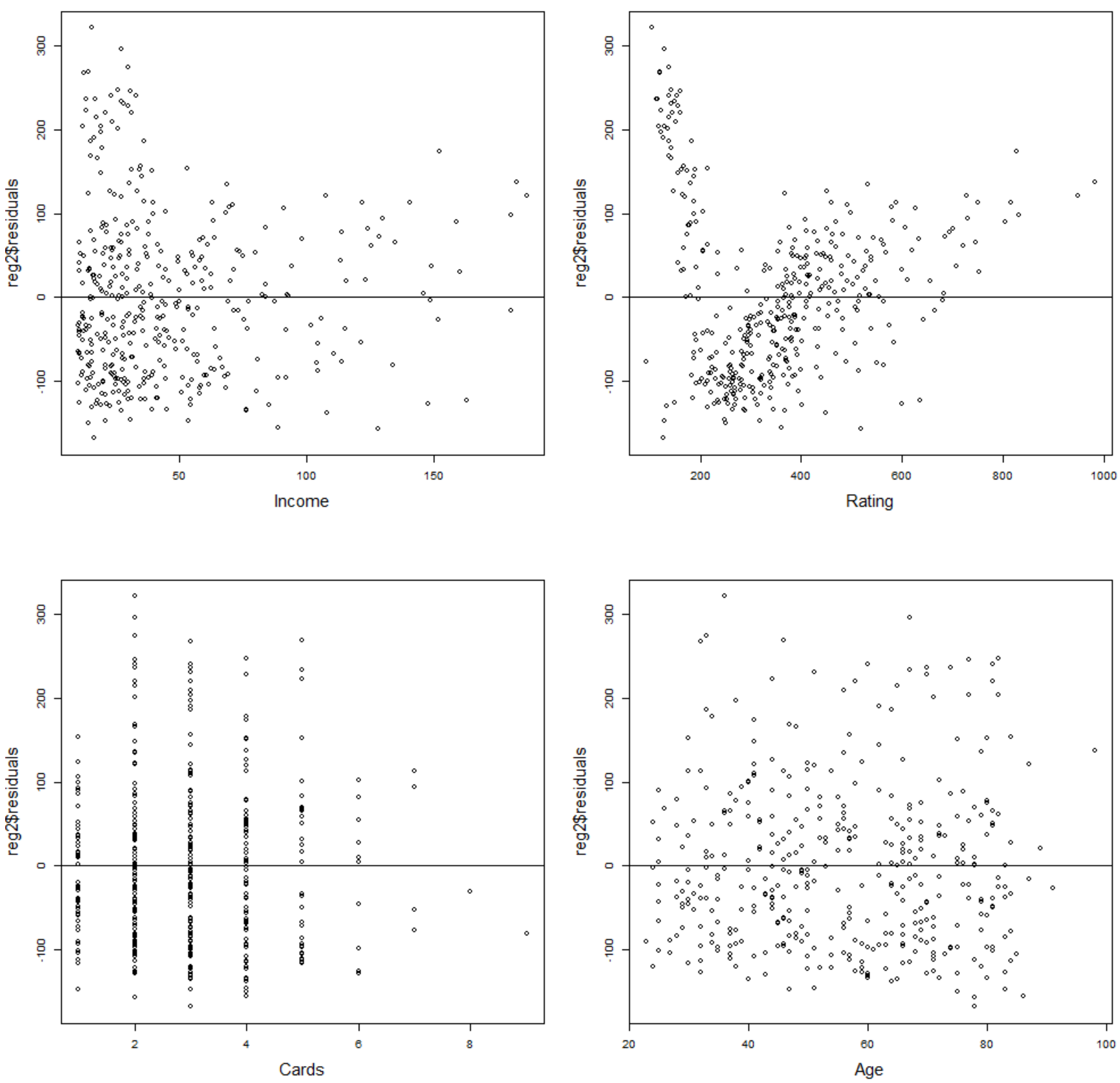


Table 4: Model Selection

	Dependent variable:		
	Balance		
	(1)	(2)	(3)
Income	−7.816***	−6.283***	−6.238***
	(0.235)	(0.488)	(0.486)
Limit	0.189***		
	(0.033)		
I(Incomeˆ2)		−0.021***	−0.021***
		(0.003)	(0.003)
Rating	1.166**	2.481***	2.471***
	(0.491)	(0.136)	(0.136)
I(Ratingˆ2)		0.002***	0.002***
		(0.0002)	(0.0002)
Cards	17.760***	2.825	
	(4.346)	(3.255)	
Age	−0.598**	−0.714***	−0.729***
	(0.295)	(0.263)	(0.261)
Edu_BinsBachelors	6.884	5.757	
	(11.901)	(10.604)	
Edu_BinsPost-Grad	−5.565	−9.563	
	(12.329)	(11.000)	
GenderFemale	−10.894	−9.118	
	(9.925)	(8.845)	
StudentYes	426.109***	429.092***	428.341***
	(16.774)	(14.920)	(14.755)
MarriedYes	−8.535	−12.380	
	(10.361)	(9.195)	
EthnicityAsian	15.965	20.348	
	(14.157)	(12.593)	
EthnicityCaucasian	10.027	13.903	
	(12.217)	(10.902)	
Constant	−497.360***	−339.084***	−329.576***
	(29.040)	(30.994)	(26.542)
Observations	400	400	400
R^2	0.955	0.964	0.964
Adjusted R^2	0.954	0.963	0.963
Residual Std. Error	98.853 (df = 387)	88.081 (df = 386)	88.218 (df = 393)
F Statistic	686.991*** (df = 12; 387)	806.545*** (df = 13; 386)	1,740.723*** (df = 6; 393)
Note:			*p<0.1; **p<0.05; ***p<0.01

Question 4

There are some limitations based on our analysis:

The linear model cannot perfectly answer this question.

Who is the fraud for using credit card?

In our model, Linear regressions are meant to describe linear relationships between variables. So, if there is a nonlinear relationship (student, Ethnicity,), then you will find limitations on it. However, we can sometimes do it by transforming some of the parameters with a log or squared.

Linear regressions are sensitive to outliers. We have lots of outliers in raw data and that's why scatter plot cannot be good fitted. I think if we can get 10X larger size then we can get a better one.

It is easy to overfit a linear model. When we have too many parameters compared to only 400 observations at model 1 then we are going to remove them one by one to find a more suitable one.

Bank's Marketing Success

Question 1

$$Pr(y_i = 1|duration_i) = \frac{\exp(\beta_0 + \beta_1 dist_i)}{1 + \exp(\beta_0 + \beta_1 dist_i)}$$
$$Pr(y_i = 0|duration_i) = \frac{1}{1 + \exp(\beta_0 + \beta_1 dist_i)}$$

Question 2

$$\ell(\beta|\mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \left[y_i \ln\left(\frac{\exp(\beta_0 + \beta_1 dist_i)}{1 + \exp(\beta_0 + \beta_1 dist_i)}\right) + (1 - y_i) \ln\left(\frac{1}{1 + \exp(\beta_0 + \beta_1 dist_i)}\right)\right]$$

Question 3

The code used is provided below

```
## Week 2
## Loading in the data
bank <- read.csv("bankTD.csv", header = TRUE, sep=",")

## Fitting the logistic model using MLE.
## The likelihood function definition.
LL_logistic<-function(beta0,beta1){
  xb=beta0+beta1*bank$duration
  lpy=bank$y*log(exp(xb)/(1+exp(xb)))+(1-bank$y)*log(1/(1+exp(xb)))
  -sum(lpy)
}

## Implementing the MLE
```

```
mle_logistic <- mle(minuslogl = LL_logistic,
                    start = list(beta0=0, beta1=0), method = "BFGS")
summary(mle_logistic)

## Using the glm function
fit1 <- glm(y~duration,binomial(link="logit"),data=bank)
summary(fit1)
```

Table 5:

	Dependent variable:	
	y	
	(mle)	(glm)
duration	0.004	0.004***
		(0.0001)
Constant	−3.180	−3.177***
		(0.067)
Observations		6,783
Log Likelihood		−2,084.980
Akaike Inf. Crit.		4,173.960
Note:	*p<0.1; **p<0.05; ***p<0.01	

Note that the mle() function doesn't provide statistical significance, standard errors, observations, a log likelihood or an AIC in its output.

Question 4

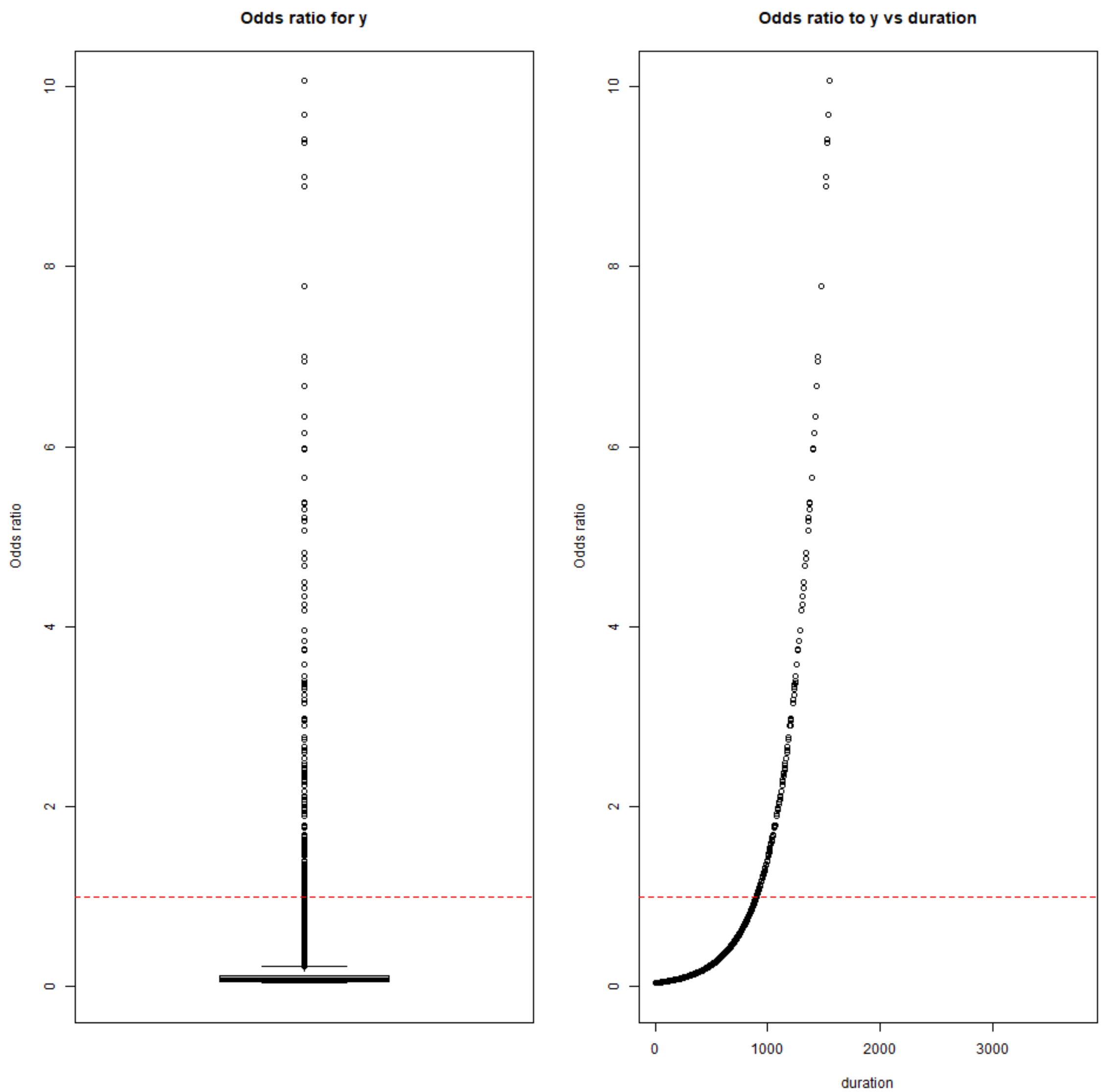
The results derived by fitting the logistic model show a positive relationship between the duration of the phone call and the likelihood that a customer takes up the term deposit product offer. Also, the parameters are statistically significant. The model implies the probabilities:

If one spends 500 seconds on the phone ($duration_i = 500$):

$$Pr(Y = 1|duration_i = 500) = \frac{\exp(-3.1768 + 0.00354 \times 500)}{1 + \exp(-3.1768 + 0.00354 \times 500)} = 0.590$$

If someone spends 1000 seconds on the phone ($duration_i = 1000$):

$$Pr(Y = 1|duration_i = 1000) = \frac{\exp(-3.1768 + 0.00354 \times 1000)}{1 + \exp(-3.1768 + 0.00354 \times 1000)} = 0.197$$



Since β_1 is positive (0.00354), the odds ratio is going to increase with the duration. The below figure shows a similar tendency as predicted. Initially, we conduct odds ratio among all data, the extreme values make a vague result. Thus, we limit the range of odds ratio to (0,10) to have a clear result. The red line (Odds ratio = 1.0) is a reference indicating that the probability of subscribing to a term deposit is equal to the probability of not subscribing. Individuals with longer duration calls (more than about 900 seconds) are more likely to sign up for the term deposit.

$$Pr(Y = 1|duration_i = 900) = \frac{\exp(-3.1768 + 0.00354 \times 900)}{1 + \exp(-3.1768 + 0.00354 \times 900)} = 0.502$$

$$Pr(Y = 0|duration_i = 900) = \frac{1}{1 + \exp(-3.1768 + 0.00354 \times 900)} = 0.498$$

$$Odds_i = \frac{Pr(y_i = 1|X_i)}{Pr(y_i = 0|X_i)} = \frac{0.502}{0.498} = 1.008$$

Also, positive β_1 implies that an increase in the duration of the call has positive impact on the propensity. The marginal effect can be calculated from:

$$ME_i = \frac{\beta_1 \exp(X_i\beta)}{(1+\exp(X_i\beta))^2}$$

In this part, we choose a couple of data points to illustrate the marginal effects.

If customers spend 255 seconds (mean) on phone call:

$$ME_i = \frac{0.00354 \exp(-3.1768 + 0.00354 \times 255)}{(1+\exp(-3.1768 + 0.00354 \times 255))^2} = 0.0003$$

If customers spend 900 seconds (mean) on phone call:

$$ME_i = \frac{0.00354 \exp(-3.1768 + 0.00354 \times 900)}{(1+\exp(-3.1768 + 0.00354 \times 900))^2} = 0.0009$$

As calculated, with the increase of duration, the magnitude is larger indicating that customers' propensity to take up the term deposit is increasing.

Question 5

Building on the model created in Section 2.3, other variables including job type, marital status, education, credit default, balance, housing, personal loan, previous contact and previous outcome information were included in the model as predictors. Using R, the following coefficients and significances were estimated.

Table 6: Multiple Logistic Regression

	Dependent variable:
	y
age	-0.001
	(0.005)
jobadmin.	0.228
	(0.266)
jobblue-collar	-0.428
	(0.264)
jobentrepreneur	-1.157^{***}
	(0.419)
jobhousemaid	-0.357
	(0.375)
jobmanagement	-0.167
	(0.255)

jobretired	0.513*
	(0.300)
jobself-employed	−0.161
	(0.333)
jobservices	−0.092
	(0.286)
jobstudent	0.902***
	(0.337)
jobtechnician	−0.305
	(0.258)
jobunknown	−1.076
	(0.804)
maritalmarried	−0.118
	(0.145)
maritalsingle	0.105
	(0.165)
educationsecondary	−0.006
	(0.159)
educationtertiary	0.494***
	(0.186)
educationunknown	0.179
	(0.261)
defaultyes	−0.704
	(0.498)
balance	0.00003***
	(0.00001)
housingyes	−0.860***
	(0.099)
loanyes	−0.504***
	(0.150)
duration	0.004***
	(0.0002)

previous	0.047**
	(0.022)
poutcomeother	0.052
	(0.228)
poutcomesuccess	2.263***
	(0.197)
poutcomeunknown	−0.534***
	(0.153)
Constant	−2.736***
	(0.424)
Observations	6,783
Log Likelihood	−1,730.257
Akaike Inf. Crit.	3,514.514
Note:	*p<0.1; **p<0.05; ***p<0.01

Based on these results, age, marital and credit default were not significant in the model. Conversely, call duration and whether the customer signed up to the previous marketing campaign had a significant and positive effect on whether the customer signed up to the term deposit. Additional predictors, such as, whether the customer had a housing or personal loan with the bank had a significant negative effect i.e. if the customer had another home or personal loan with the bank, they were less likely to open a further term deposit.

Predictors within the job and education categories had mixed results. Customers with tertiary education were significantly more likely to sign up to the term deposit compared to customers with only primary education (reference group). However, customers with secondary or unknown education were not significantly different from those with primary education only. Within the jobs category, customers that were within blue collar, technician, entrepreneur and management type roles were significantly less likely than those in admin jobs (reference group) to sign up for the term deposit. In contrast, students were significantly morel likely to open a term deposit than customers in admin jobs. All other jobs were not deemed significantly different from admin roles.

To understand if the multiple predictor model improved upon the single predictor model from Section 2.3, the AIC values were compared. The table below summarises the deviance and AIC values for the null, single parameter and multiple parameter models. The AIC value was significantly reduced in the multiple predictor model compared to the single predictor model. This indicates that the multiple predictor model is a better predictor of whether customer will sign up for the term deposit than the single predictor model, hence is an improvement upon the single predictor model. Both multiple and single predictor models perform better than the null model based on the AIC.

Question 6

Table 7: Model Selection

	Dependent variable:		
	y		
	(1)	(2)	(3)
age	−0.001	0.0003	
	(0.005)	(0.006)	
jobadmin.	0.228	0.340	0.348
	(0.266)	(0.273)	(0.272)
jobblue-collar	−0.428	−0.328	−0.356
	(0.264)	(0.269)	(0.266)
jobentrepreneur	−1.157***	−1.065***	−1.101***
	(0.419)	(0.413)	(0.412)
jobhousemaid	−0.357	−0.246	−0.265
	(0.375)	(0.380)	(0.376)
jobmanagement	−0.167	−0.085	−0.113
	(0.255)	(0.260)	(0.259)
jobretired	0.513*	0.579*	0.528*
	(0.300)	(0.307)	(0.289)
jobself-employed	−0.161	−0.002	−0.019
	(0.333)	(0.337)	(0.335)
jobservices	−0.092	0.067	0.062
	(0.286)	(0.290)	(0.289)
jobstudent	0.902***	1.115***	1.229***
	(0.337)	(0.348)	(0.336)
jobtechnician	−0.305	−0.180	−0.185
	(0.258)	(0.264)	(0.262)
jobunknown	−1.076	−1.164	−1.208
	(0.804)	(0.822)	(0.824)
maritalmarried	−0.118	−0.094	
	(0.145)	(0.149)	
maritalsingle	0.105	0.139	

	(0.165)	(0.168)	
educationsecondary	−0.006	−0.004	0.015
	(0.159)	(0.161)	(0.160)
educationtertiary	0.494***	0.524***	0.569***
	(0.186)	(0.189)	(0.186)
educationunknown	0.179	0.170	0.186
	(0.261)	(0.267)	(0.267)
defaultyes	−0.704	−0.774	−0.720
	(0.498)	(0.510)	(0.507)
balance	0.00003***	0.00005**	0.00004***
	(0.00001)	(0.00002)	(0.00001)
I(balanceˆ2)		−0.000	
		(0.000)	
housingyes	−0.860***	−0.972***	−0.974***
	(0.099)	(0.107)	(0.106)
loanyes	−0.504***	−0.458***	−0.507***
	(0.150)	(0.158)	(0.150)
duration	0.004***	0.007***	0.007***
	(0.0002)	(0.0004)	(0.0004)
I(durationˆ2)		−0.00000***	−0.00000***
		(0.00000)	(0.00000)
previous	0.047**	−0.014	
	(0.022)	(0.040)	
poutcomeother	0.052	0.096	0.089
	(0.228)	(0.237)	(0.237)
poutcomesuccess	2.263***	2.408***	2.401***
	(0.197)	(0.206)	(0.206)
poutcomeunknown	−0.534***	−0.577***	−0.593***
	(0.153)	(0.161)	(0.159)
housingno:previous			−0.022
			(0.039)
housingyes:previous		0.093**	0.069***

		(0.043)	(0.026)
loanyes:previous		−0.058	
		(0.068)	
Constant	−2.736***	−3.566***	−3.543***
	(0.424)	(0.443)	(0.323)
Observations	6,783	6,783	6,783
Log Likelihood	−1,730.257	−1,652.637	−1,655.544
Akaike Inf. Crit.	3,514.514	3,367.273	3,363.087
Note:			*p<0.1; **p<0.05; ***p<0.01

Question 7

The cut-off point in the logistic regression model represents the border of true or false predictions. Given the regression model explained before with 50% cut-off rate, the precision of the model is 64.1%, the error rate with false negative is 8.07%, and the error rate of false positive is 35.9%, as shown in the following table.

Table 8: Hit and Miss Results

	Precision	False-Negatives	False-Positives
0.05	0.241	0.018	0.759
0.1	0.345	0.029	0.655
0.15	0.417	0.041	0.583
0.2	0.474	0.048	0.526
0.25	0.510	0.056	0.490
0.3	0.551	0.060	0.449
0.35	0.580	0.065	0.420
0.4	0.606	0.071	0.394
0.45	0.635	0.074	0.365
0.5	0.641	0.081	0.359
0.55	0.676	0.084	0.324
0.6	0.694	0.090	0.306
0.65	0.709	0.094	0.291
0.7	0.727	0.099	0.273
0.75	0.767	0.103	0.233
0.8	0.778	0.106	0.222
0.85	0.800	0.110	0.200
0.9	0.825	0.113	0.175
0.95	0.875	0.115	0.125

Since the cut-off rate determines the sensitivity of the model prediction, which is a trade-off between false negative and false positive, the false-negative rate will increase simultaneously with the cut-off rate, and the false-negative rate is the opposite. The precision rate will reach the peak with the cut-off rate reaches 0.45-0.5, as shown in the following table.

The recommended cut-off rate for this model is 0.45. This model's precision reaches its peak, and the false-negative rate is also lower than choosing the 0.5 cut-off rate. It is clear that the prediction model with a 0.45 cut-off rate is more balanced and performs better.