MATH-661: Generalized Linear Models
Midterm Exam – Sample Solution

Part I: Aspirin and heart attack [20 points]

A study investigating the association between heart attacks and the use of aspirin is conducted. Age is a potential confounder and is also considered. The following indicator variables are defined:

$$Y = \begin{cases} 1 & \text{if heart attack} \\ 0 & \text{if no heart attack} \end{cases} \qquad \text{Aspirin} = \begin{cases} 1 & \text{if aspirin} \\ 0 & \text{if placebo} \end{cases}$$

$$\text{Age1} = \begin{cases} 1 & \text{if age is } 40-50 \\ 0 & \text{otherwise} \end{cases} \qquad \text{Age2} = \begin{cases} 1 & \text{if age is } >50 \\ 0 & \text{otherwise} \end{cases}$$

The following table shows the results of fitting logistic regression models for $P(Y=1)$:

Model	Covariates	Estimate $\hat{\beta}$	Standard Error	log-likelihood
1	None	-2.99	0.19	-116.54
2	Aspirin	-0.82	0.41	-114.41
3	Age1	-0.19	0.47	-116.27
	Age2	0.17	0.45	
4	Aspirin	-0.82	0.41	-114.14
	Age1	-0.18	0.47	
	Age2	0.19	0.45	
5	Aspirin	-0.65	0.63	-113.83
	Age1	-0.22	0.59	
	Age2	0.39	0.54	
	(Age1)*Aspirin	0.10	0.97	
	(Age2)*Aspirin	-0.68	1.03	

(a) **Test the null hypothesis of constant aspirin effect on the risk of heart attack across age groups (i.e., no interaction between aspirin and age).**

$$H_0: \text{ no interaction between aspirin and age} \qquad vs. \qquad H_1: \text{ there is an interaction}$$

This is equivalent to comparing the restricted model with additive effects for age and aspirin (Model 4) to the saturated model with interaction terms (Model 5). The likelihood ratio test is given by

$$\Delta = -2(\log L_4 - \log L_5) = -2(-114.14 + 113.83) = 0.62$$

and has $df = 6 - 4 = 2$. So the p-value $= P(\chi^2_2 > 0.62) = 0.733$.

Therefore, we fail to reject H_0. There is no evidence of a varying effect of aspirin across age gruops.

(b) **Based on the model with additive/main effects for age and aspirin (Model 4)**

i. **Calculate the MLE of the odds ratio of aspirin use on heart attack, adjusting for age. Provide a 95% confidence interval for this odds ratio and interpret it in context.**

Adjusting for age, the estimated effect of aspirin is

$$\hat{\beta}_1 = \log \widehat{OR} = -0.82$$

therefore, the MLE of the odds ratio is

$$\widehat{OR} = \exp(-0.82) = 0.44$$

The odds of heart attack for subjects taking aspirin is 0.44 times lower than for those taking aspirin, controlling for age. In other words, adjusting for age, the odds of heart attack are $1/0.44 = 2.27$ times higher for people not taking aspiring compared to those taking aspirin.
A 95% confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm z_{0.025}\ SE(\hat{\beta}_1) = -0.82 \pm 1.96 \times 0.41 = (-1.624, -0.016)$$

and an 95% CI for the OR is

$$\exp\left(\hat{\beta}_1 \pm z_{0.025}\ SE(\hat{\beta}_1)\right) = (e^{-1.624}, e^{-0.016}) = (0.197, 0.984)$$

The odds of heart attack are between 0.2 times and 0.98 times lower among subjects taking aspirin, adjusting for age.

ii. **Perform a Wald test of the null hypothesis that there is no effect of aspirin on the risk of heart attack, controlling for age. What do you conclude?**

$$H_0 : \beta_1 = 0 \qquad vs. H_1 : \beta_1 \neq 0$$

The Wald test for this hypothesis is given by

$$\frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{-0.82}{0.41} = 2$$

and has p-value $= 2 \times P(Z > 2) = 0.0455$. Therefore, we reject H_0 at $\alpha = 0.05$. There is evidence of an association between aspirin use and heart attack, controlling for age.

iii. **Perform a likelihood ratio test of the null hypothesis that there is no effect of age on the risk of heart attack, controlling for aspirin use. State your conclusion.**

We now want to test

$$H_0: \text{ no age effect} \qquad vs. \qquad H_1: \text{ there is an age effect}$$

This can be evaluated using a likelihood ratio test comparing Model 2 (aspirin effect only) to Model 4 (age+aspirin effects)

$$\Delta = -2(\log L_2 - \log L_4) = -2(-114.41 + 114.14) = 0.54 \qquad df = 4 - 2 = 2$$

so p-value $= P(\chi^2_2 > 0.54) = 0.76$.

We fail to reject H_0. There is no evidence of an association between age group and heart attack, given treatment.

(c) **Evaluate the deviance of each model provided in the table and assess its goodness-of fit. Which models do not provide adequate fit to the data?**

We are testing

$$H_0: \text{ model fits data well} \qquad vs \qquad H_1: \text{ model does not fit the data well}$$

Let us collect the deviance statistics from all the models, which is given by

$$G^2 = -2(\log L_M - \log L_S) \sim \chi^2_{df}$$

where L_S is the log-likelihood of the saturated model and L_M is the log-likelihood of the model under consideration

Model	G^2	df	p-value
(5) saturated	0.00	0	–
(4) aspirin + age	0.62	2	0.733
(3) age	4.88	3	0.181
(2) aspirin	1.16	4	0.885
(1) null	5.42	5	0.367

For all the models considered, the goodness-of-fit p-value is large, so there's no evidence that the models do not fit the data well. The model with only age has a relatively low goodness-of-fit p-value = 0.18, which makes its fit to the data questionable.

(d) **Perform model selection using analysis-of-deviance. Make sure you describe all the steps to arrive at your final model.**

Now, we are testing nested models using the change in deviance, which is equivalent to doing the likelihood ratio test:

H_0 : smaller model fits data as well as larger model vs H_1 : smaller model does not fit the data as well as l

Models compared	ΔG^2	Δdf	p-value
null vs. aspirin	4.26	1	0.039
null vs. age	0.54	2	0.763
age vs. (aspirin+age)	4.26	1	0.039
aspirin vs. (aspirin+age)	0.54	2	0.763
(aspirin+age) vs saturated	0.62	2	0.733

We note that we fail to reject H_0 for the comparisons of the null vs the model with age only, aspirin vs. (aspirin+age), and (aspirin+age) vs. saturated. The model with aspirin fits significantly better than the null model; and the model with aspirin fits as well as the model with (aspirin+age). Thus, the most parsimonious model that provides the best fit to the data is the model with aspirin.

Part II: Credit risks for bank loan [30 points]

Banks want to reduce the rate of loan defaults. Loan officers want to be able to identify characteristics that are indicative of people who are likely to default on loans, and then use those characteristics to identify good and bad credit risks.

Financial and demographic information are collected on 850 past and prospective customers. Of these, 700 are customers who were previously given loans and 150 are prospective customers that the bank needs to classify as good of bad credit risks. The data are saved in `Bank_loan.txt` and contain the following variables:

age	age in years
ed	highest level of education
	1: did not complete high school; 2: high school degree
	3: some college; 4: college degree; 5: post-bachelor degree
employ	years with current employer
address	years at current address
income	household income in thousands
debtinc	debt to income ratio ($\times 100$)
creddebt	credit card debt in thousands
othdebt	other debt in thousands
default	previously defaulted – 0: No, 1: Yes, NA: prospective customers

1. **Exploratory data analysis & data processing**

   (a) **Provide appropriate summary statistics and graphical displays for the variables in the data. Discuss their distributions.**

   Figure 1 shows histograms and boxplots for the continuous variables, and bar plots for the categorical variables. Tables 1 and 2 provide summary statistics for the continuous and categorical variables, respectively. We note that age has a fairly symmetric distribution with a slight skewness to the right with a mean of 35.03 (median = 34) and a standard deviation of 8.04 (IQR = 12). The other continuous variables are all skewed to the right:

   - the median for years with current employer is 7 with IQR = 10
   - the median for years at current address is 7 with IQR = 9
   - the median income is \$35,000 with IQR = \$31.75
   - the median debt to income ratio is 8.7% with IQR = 8.7%
   - the median credit card debt is \$885 with IQR = \$1,516
   - the median other debt is \$2,003 with IQR = \$2,857

   For highest education level, we see that a majority of the customers did not complete high school (54.1%), about a quarter (27.6%) have a high school degree, 11.9% have some college experience,

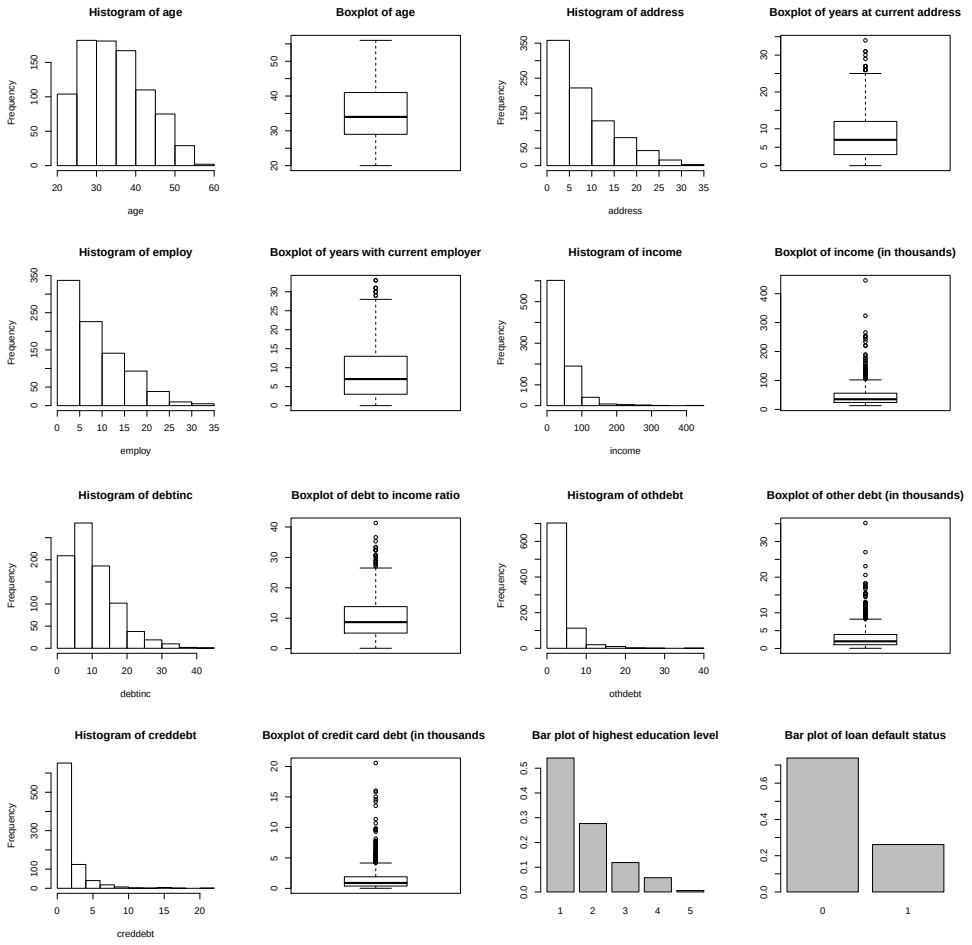

Figure 1: Graphical display of variables in the data

only 5.8% have a college degree and 0.6% have a post-bachelor degree. About a quarter (26.1%) of the 700 past customers have defaulted on their loans.

```
par(mfrow=c(2,2))
hist(age); boxplot(age, main="Boxplot of age")
hist(employ); boxplot(employ, main="Boxplot of years with current employer")
hist(address); boxplot(address, main="Boxplot of years at current address")
hist(income); boxplot(income, main="Boxplot of income (in thousands)")
hist(debtinc); boxplot(debtinc, main="Boxplot of debt to income ratio")
hist(creddebt); boxplot(creddebt, main="Boxplot of credit card debt (in thousands)")
hist(othdebt); boxplot(othdebt, main="Boxplot of other debt (in thousands)")
barplot(prop.table(table(ed)), main="Bar plot of highest education level")
barplot(prop.table(table(default)), main="Bar plot of loan default status")
summary(age); sd(age)
summary(employ); summary(address)
```

	Minimum	Q_1	Median	Q_3	Maximum	IQR
age	20	29	34	41	56	12
			(mean = 35.03)			(sd = 8.04)
employ	0	3	7	13	33	10
address	0	3	7	12	34	9
income	13	24	35	55.75	446	31.75
debtinc	0.10	5.10	8.70	13.80	41.30	8.70
creddebt	0.0117	0.3822	0.8851	1.8984	20.5613	1.5162
othdebt	0.04558	1.04594	2.00324	3.90300	35.19750	2.85706

Table 1: Summary statistics for continuous variables

Variable	levels	n	%
education	did not complete high school	460	54.12%
	high school degree	235	27.65%
	some college	101	11.88%
	college degree	49	5.76%
	post-bachelor degree	5	0.59%
loan default	Yes	183	26.14%
	No	517	73.86%

Table 2: Summary statistics for categorical variables

```
summary(income); summary(debtinc)
summary(creddebt); summary(othdebt)

table(ed); prop.table(table(ed))
table(default); prop.table(table(default))
```

(b) **Since there are few observations with post-bachelor degree (`ed` = 5), combine these with the group with college degree (`ed` = 4). You will be using education with these 4 levels in subsequent analyses.**

We create a new variable `edu` combining `ed=4` and `ed=5`:

```
Bank_loan$educ = Bank_loan$ed
Bank_loan$educ[Bank_loan$ed==5] = 4

> table(Bank_loan$educ)
  1   2   3   4
460 235 101  54
```

(c) **Separate the 150 prospective customers for whom credit risk is to be predicted from the 700 past customers.**

The prospective customers all have NA for default loan, so it is easy to subset them using this information:

```
past = Bank_loan[!is.na(default),]
```

2. **Model building & diagnostics** – use the 700 past customers for this task.

(a) **Perform stepwise selection.**

The stepwise selection chooses the model with debtinc, employ, creddebt, address and age:

```
fitpast.all = glm(default ~ as.factor(educ)+age+employ+address+income
           +debtinc+creddebt+othdebt, family=binomial, data=past)
fitpast.null = glm(default ~ 1, family=binomial, data=past)

> step(fitpast.null, scope=list(lower=fitpast.null, upper=fitpast.all))
Start:  AIC=806.36
default ~ 1

                  Df Deviance    AIC
+ debtinc          1   701.43 705.43
```

```
+ employ           1   739.42 743.42
+ creddebt         1   765.86 769.86
+ address          1   783.91 787.91
+ age              1   790.73 794.73
+ othdebt          1   790.76 794.76
+ as.factor(educ)  3   793.63 801.63
+ income           1   800.37 804.37
<none>                 804.36 806.36

Step:  AIC=705.43
default ~ debtinc

                  Df Deviance    AIC
+ employ           1   631.08 637.08
+ address          1   675.79 681.79
+ age              1   684.45 690.45
+ as.factor(educ)  3   689.02 699.02
+ othdebt          1   694.27 700.27
+ income           1   697.75 703.75
+ creddebt         1   699.15 705.15
<none>                 701.43 705.43
- debtinc          1   804.36 806.36

Step:  AIC=637.08
default ~ debtinc + employ

                  Df Deviance    AIC
+ creddebt         1   575.64 583.64
+ income           1   615.45 623.45
+ address          1   622.35 630.35
+ othdebt          1   626.59 634.59
<none>                 631.08 637.08
+ as.factor(educ)  3   625.25 637.25
+ age              1   631.08 639.08
- employ           1   701.43 705.43
- debtinc          1   739.42 743.42

Step:  AIC=583.64
default ~ debtinc + employ + creddebt
```

```
                  Df Deviance    AIC
+ address          1   556.73 566.73
<none>                 575.64 583.64
+ income           1   574.87 584.87
+ age              1   575.00 585.00
+ othdebt          1   575.63 585.63
+ as.factor(educ)  3   573.79 587.79
- debtinc          1   599.42 605.42
- creddebt         1   631.08 637.08
- employ           1   699.15 705.15

Step:  AIC=566.73
default ~ debtinc + employ + creddebt + address

                  Df Deviance    AIC
+ age              1   553.18 565.18
<none>                 556.73 566.73
+ othdebt          1   556.42 568.42
+ income           1   556.72 568.72
+ as.factor(educ)  3   554.89 570.89
- address          1   575.64 583.64
- debtinc          1   580.01 588.01
- creddebt         1   622.35 630.35
- employ           1   667.22 675.22

Step:  AIC=565.18
default ~ debtinc + employ + creddebt + address + age

                  Df Deviance    AIC
<none>                 553.18 565.18
- age              1   556.73 566.73
+ income           1   553.02 567.02
+ othdebt          1   553.02 567.02
+ as.factor(educ)  3   550.86 568.86
- address          1   575.00 585.00
- debtinc          1   576.98 586.98
- creddebt         1   618.56 628.56
- employ           1   661.59 671.59

Call:  glm(formula = default ~ debtinc + employ + creddebt + address +
```

```
    age, family = binomial, data = past)

Coefficients:
(Intercept)      debtinc       employ     creddebt      address          age
   -1.63128      0.08926     -0.26076      0.57265     -0.10365      0.03256

Degrees of Freedom: 699 Total (i.e. Null);  694 Residual
Null Deviance:     804.4
Residual Deviance: 553.2  AIC: 565.2

fit.step = glm(default ~ age+employ+address+debtinc+creddebt, family=binomial, data=past)
> summary(fit.step)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.63128    0.51268  -3.182  0.00146 **
age          0.03256    0.01717   1.896  0.05799 .
employ      -0.26076    0.03011  -8.662  < 2e-16 ***
address     -0.10365    0.02309  -4.490 7.13e-06 ***
debtinc      0.08926    0.01855   4.813 1.49e-06 ***
creddebt     0.57265    0.08723   6.565 5.20e-11 ***
---

    Null deviance: 804.36  on 699  degrees of freedom
Residual deviance: 553.18  on 694  degrees of freedom
AIC: 565.18
```

i. **Provide the equation of the selected model.**

The equation of the stepwise selected model is:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -1.63 + 0.03\ Age - 0.26\ Employ - 0.10\ Address + 0.09\ Debtinc + 0.57\ Creddebt$$

ii. **Interpret the effect of each of the covariates in the selected model.**
Age is marginally significant (p-value = 0.058) and all the other variables in the model are statistically significant, so we can interpret their regression coefficients.

- Adjusting for years with current employer, years at current address, debt to income ratio and credit card debt, for each additional year in age, the log-odds of defaulting on a loan increases by 0.03.
- Controlling for age, years at current address, debt to income ratio and credit card debt, for each additional year with current employer, the log-odds of defaulting on a loan decreases

by 0.26.

* Holding age, years with current employer, debt to income ratio and credit card debt constant, for each additional year at current address, the log-odds of defaulting on a loan is -0.10.
* Adjusting for age, years with current employer, years at current address and credit card debt, for each additional percent in debt to income ratio, the odds ratio of defaulting on a loan is $\exp(0.089) = 1.09$ times higher.
* Controlling for age, years with current employer, years at current address and debt to income ratio, for each additional \$1,000 in credit card debt, the odds ratio of defaulting on a loan is $\exp(0.57265) = 1.77$.

Thus, customers who have spent less time at either their current employer or their current address, as well as customers with higher debt-to-income ratios or higher amounts of credit card debt are at higher risk defaulting on their loans.

iii. **Assess the goodness-of-fit of the selected model.**

Since we have ungrouped data with no/few replicates by covariate pattern, we cannot use the deviance to assess the model goodness-of-fit. We resort to the Hosmer-Lemeshow goodness of fit test:

$$H_0: \text{ model fits the data well} \qquad vs \qquad H_1: \text{ model does not fit data well}$$

which leads to p-value of 0.8879. Thus, we fail to reject H_0. The model appears to provide an adequate fit to the data.

```
library(ResourceSelection)
gof = hoslem.test(past$default, fitted(fit.step))
```

```
> gof
Hosmer and Lemeshow goodness of fit (GOF) test

data:  past$default, fitted(fit.step)
X-squared = 3.6418, df = 8, p-value = 0.8879
```

(b) **Perform a lasso variable selection using the misclassification error as criterion for choosing λ.**

```
X = model.matrix(default ~ as.factor(educ)+age+employ+address+income+debtinc+creddebt+othdeb
Y = past$default
cvfit = cv.glmnet(x=X[,-1], y=Y, family="binomial", type.measure="class")
```

i. **Compare the models selected using `lambda.1se` to the stepwise selected model in (a). Perform a likelihood ratio test to choose the preferred model between the two at $\alpha = 0.05$.**

```
> coef(cvfit, s=cvfit$lambda.1se)
11 x 1 sparse Matrix of class "dgCMatrix"
                            1
(Intercept)      -1.17536120
as.factor(educ)2  .
as.factor(educ)3  .
as.factor(educ)4  .
age               .
employ           -0.12356767
address          -0.02420107
income            .
debtinc           0.07469154
creddebt          0.23738602
othdebt            .
fit.lasso = glm(default ~ employ+address+debtinc+creddebt, family="binomial", data=past)
> summary(fit.lasso)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.79107    0.25154  -3.145  0.00166 **
employ      -0.24260    0.02806  -8.646  < 2e-16 ***
address     -0.08125    0.01960  -4.145 3.39e-05 ***
debtinc      0.08827    0.01854   4.760 1.93e-06 ***
creddebt     0.57300    0.08727   6.566 5.18e-11 ***
---
    Null deviance: 804.36  on 699  degrees of freedom
Residual deviance: 556.73  on 695  degrees of freedom
AIC: 566.73
```

The model based on `lambda.1se` selects employ, address, debtinc and creddebt. This model is nested in the model chosen by stepwise selection, which contains these same variables along with age. The likelihood ratio comparing these two nested models can be obtained by considering their difference in deviance:

$$-2(\log L_{lasso} - \log L_{step}) = G^2_{lasso} - G^2_{step} = 556.7317 - 553.1761 = 3.556$$

which follows a chi-square distribution with 1 *df* and leads to a p-value = 0.06.

```
> deviance(fit.lasso)-deviance(fit.step)
[1] 3.555619
> 1-pchisq(deviance(fit.lasso)-deviance(fit.step), 1)
[1] 0.05934418
```

We fail to reject H_0 at $\alpha = 0.05$. Thus, we would choose the model selected with lasso.

(c) For the model selected based on lasso

i. **Identify observations with unusual/outlying standardized residuals. How do the predictions for these individuals, based on their fitted values $\hat{\pi}_i$, compare to their observed default status?**

The standardized residuals follow a $N(0,1)$ distribution and are given by

$$r_i = \frac{e_i}{\sqrt{1 - h_{ii}}}$$

Thus value outside of the $(-3, 3)$ range would be considered outliers. There are 13 observations that are deemed to be outliers. Their fitted values, $\hat{\pi}_i$, and the observed loan default status are given below. We note that the estimated probability of defaulting for observation based on the lasso-selected model is 0.95, so this customer is deemed highly likely to default, but he did not default. The remaining 12 outliers are observations with estimated probabilities of defaulting less than 0.10 (some are as low as 0.017), so deemed highly unlikely to default based on the lasso-selected model, but in fact ended up defaulting on their loans.

```
stdres = resid(fit.lasso, type="pearson")/sqrt(1-hatvalues(fit.lasso))
outres = which(abs(stdres) > 3)
```

```
> cbind(fit.lasso$fitted.values[outres], past$default[outres], stdres[outres])
          [,1] [,2]      [,3]
16  0.09670408    1  3.060687
36  0.95005352    0 -4.376107
53  0.03700212    1  5.111825
62  0.07325835    1  3.565179
107 0.09169451    1  3.156309
152 0.01725101    1  7.558462
187 0.04883231    1  4.426018
202 0.05505157    1  4.148419
219 0.07380372    1  3.547368
281 0.03470598    1  5.280064
515 0.05618503    1  4.107892
678 0.07119029    1  3.617729
696 0.05170308    1  4.290266
```

ii. **Using a cut-off of 0.3 for predicting whether a person defaults or not on a loan**

The predicted status will thus be

$$\begin{cases} \hat{y}_i = 1 & \text{if } \hat{\pi}_i > 0.3 \\ \hat{y}_i = 0 & \text{if } \hat{\pi}_i \leq 0.3 \end{cases}$$

```
yhat.past = as.numeric(fitted(fit.lasso)>0.3)
```

```
> table(past$default, yhat.past)
   yhat.past
      0   1
  0 415 102
  1  46 137
```

- **what proportion of the 700 customers would have been predicted as defaulting (and thus would have been denied a loan)?**

  $102 + 137 = 239$ customers, that is 34.1% of the customers, would have been predicted to default on their loans.

- **what would be the misclassification rate?**

  $46 + 102 = 148$ customers, that is 21.1% of the customers, would have been misclassified.

- **what would be the misclassification rate among the defaulters?**

  $46/(46 + 137) = 25.1\%$ of the defaulters would have been misclassified. That is, a quarter of the customers who ended up defaulting would have been mistakenly predicted as non-defaulters.

- **what would be the misclassification rate among the non-defaulters?**

  $102/(415 + 102) = 19.7\%$ of the non-defaulters would have been misclassified. That is, about 20% of the customers who ended up not defaulting would have been mistakenly predicted as defaulters.

iii. **Provide the ROC curve and the area under the ROC curve for the selected model.**

Figure 2 provides the ROC curve for the lasso-selected model. The area under the ROC curve is 0.856.

```
library(ROCR)
```

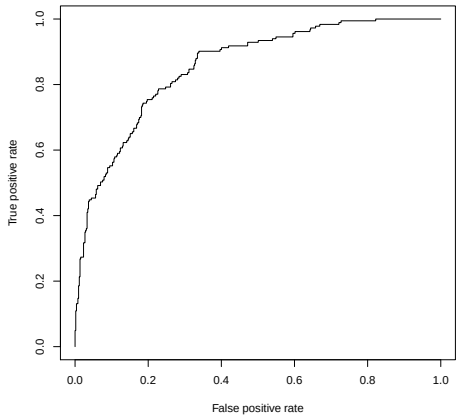

Figure 2: ROC curve for lasso-selected model

```
perf = performance(pred, "tpr", "fpr")
par(mfrow=c(1,1))
plot(perf)

> performance(pred, "auc")@y.values
[[1]]
[1] 0.8556088
```

3. **Prediction for future customers** – consider the model selected based on lasso.

   (a) **Calculate the predicted probabilities of loan default for the 150 prospective customers.** The linear predictor for the i-th customer is estimated by:

$$\hat{\eta}_i = -0.79 - 0.24\ Employ_i - 0.08\ Address_i + 0.09\ Debtinc_i + 0.57\ Creddebt_i$$

   and the predicted probabilities are given by:

$$\hat{\pi}_i = \frac{\exp(\hat{\eta}_i)}{1 + \exp(\hat{\eta}_i)}$$

```
future = Bank_loan[is.na(default),]        # subsetting data of future customers
yhat.future = predict(lasso.fit, newdata=future, type="response")
```

   (b) **Provide a histogram and a boxplot of the predicted probabilities.**

   Figure 3 shows a histogram and a boxplot of the predicted probabilities of defaulting on a loan. The distribution of the predicted probabilities is skewed to the right with a large proportion of customers having small predicted values. About a quarter of the potential customers have a predicted probability of defaulting lower than 0.03, about half have a predicted probability of

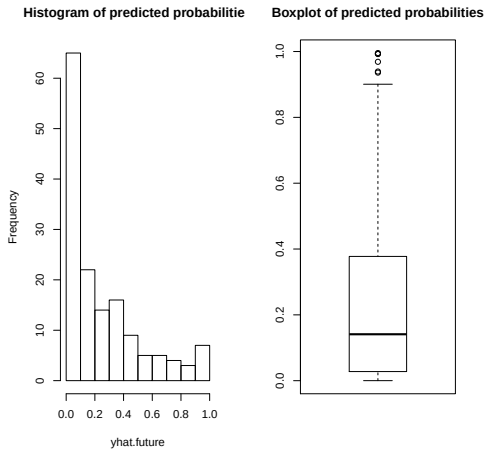

Figure 3: Histogram and boxplot of predicted probabilities of defaulting on loan for prospective customers

defaulting lower than 0.14. Less than a quarter have a predicted probability of defaulting greater than 0.4, with only 10 potential customers having a predicted probability of defaulting greater than 0.8.

```
par(mfrow=c(1,2))
hist(yhat.future, main="Histogram of predicted probabilities")
boxplot(yhat.future, main="Boxplot of predicted probabilities")
```

```
> summary(yhat.future)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0001166 0.0283204 0.1410435 0.2452652 0.3762677 0.9954647
```

(c) **Using a cut-off of 0.3, how many of the 150 prospective customers would be expected to default on a loan?**

Using a cut-off of 0.3, 49 of the 150 customers would be predicted as defaulting on their loan.

```
> table(yhat.future>0.3)

FALSE  TRUE
  101    49
```