Machine Learning in Finance Assignment 2

1) We analyze a dataset, which is customer's record from a bank about loan application. It contains 7 columns:

Column	Name	Descriotion	Value
1	Age	Age	Continuous
2	Addr	Mean time at address	Continuous
3	Empl	Mean time with employers	Continuous
4	Bank	Time with bank	Continuous
5	House	Monthly housing expense	Continuous
6	Save	Savings account balance	Continuous
7	Result	Result (1=accept, 0=reject)	Binary

Note that the last column is the target variable. We create a binary variable g = 1 if Bank< 2.5 and g = 0 otherwise. A logistic regression is fitted with the following result:

Coefficients:

	Estimate	Std.Error	z-value	$\Pr(>\|z\|)$	
(Intercept)	0.23	0.0455	5.040	3.26e-07	***
Addr	0.13	0.0611	2.125	0.0170	*
Empl	0.22	0.0412	5.334	7.29e-08	***
g	-1.82	0.3700	-4.909	6.21e-07	***
Save	0.0005	0.0001	4.102	2.39e-05	***
Addr:g	-0.11	0.0628	-1.752	0.0494	*

a) Suppose that a customer's record is as follows:

$$(Age, Addr, Empl, Bank, House, Save) = (30, 7.5, 15, 1.5, 900, 800)$$

What is the estimated probability of Result = 1 for this customer?

b) With the logistic regression fit and the Bayes rule, we construct a classifier and obtain the following confusion matrix for 1,000 test data: Compute the overall faction of correct predictions, the false positive rate, and the false negative rate.

		True status	
		Accept (-)	Reject (+)
Predicted status	Accept (-)	710	179
	Reject (+)	26	85

c) If we want to have a lower false negative rate, how should we adjust our classifier?

d) Suppose that we create another binary variable h such that h = 2 if Bank < 2.5 and h = 1 otherwise. A logistic regression is fitted with the following model:

$$\ln \pi/(1-\pi) = \beta_0 + \beta_1 \text{Addr} + \beta_2 \text{Empl} + \beta_3 \text{h} + \beta_4 \text{Save} + \beta_5 \text{h} * \text{Addr},$$

where π = P(Result = 1). Write down the maximum likelihood estimates (MLEs) of the parameters β0 to β5 in this model. Show the detail of your calculations.

Logistic regression question

a) estimated probability of Result = 1 for this customer

$$\mathrm{P}(\mathrm{Result} = 1) = \frac{1}{1 + \exp(1.59 - 0.02 * 7.5 - 0.22 * 15 - 0.0005 * 800)} = 0.9055.$$

b) False positive rate and False negative rate

	(-) Actual	(+) Actual
(-) Predicted	710	179
(+) Predicted	26	85
	736	264

$$False\ positive\ rate = \frac{26}{736} = 0.035 = 3.5\%$$

$$False\ negative\ rate = \frac{179}{236} = 0.678 = 67.8\%$$

c) The original classifier classifies result as acceptance if the estimated P(Result =1) > 0.5. If we want to have a lower false negative rate, that means less misclassification of rejected applications as accepted, we can set a higher threshold, e.g. thres = 0.7. Then the revised classifier, that classifies result as acceptance if the estimated P(Result = 1) > 0.7, will have lower false negative rate

d) Notice that h = g + 1.

$$ln\ \pi/(1 - \pi) = \beta_0 + \beta_1 Addr + \beta_2 Empl + \beta_3 h + \beta_4 Save + \beta_5 h * Addr$$
$$= \beta_0 + \beta_1 Addr + \beta_2 Empl + \beta_3 (g + 1) + \beta_4 Save + \beta_5 (g + 1) * Addr$$
$$= (\beta_0 + \beta 3) + (\beta_1 + \beta 5) Addr + \beta_2 Empl + \beta_3 g + \beta_4 Save + \beta_5 g * Addr$$

$$\beta_0 + \beta_3 = 0.23, \beta_1 + \beta_5 = 0.13, \beta_2 = 0.22, \beta_3 = -1.82, \beta_4 = 0.0005, \beta_5 = -0.11$$

$$\beta_0 = 2.05, \beta_1 = 0.24, \beta_2 = 0.22, \beta_3 = -1.82, \beta_4 = 0.0005, \beta_5 = -0.11$$

2) A scientist has measured expression levels for 5,000 genes on 180 patients. He is interested in building a regression model based upon the genes to predict a continuous clinical response and decides to use the LASSO. He found that it is necessary to standardize the gene expression values. He implements the following procedure to build his model:

1. Standardize all gene features such that each feature has zero sample mean and unit sample standard deviation, computed with 180 data. Note that the response is not standardized.
2. Employ 5-fold cross-validation on the standardized 5,000 genes to select the tuning parameter (λ) in the LASSO and report the prediction error.

a) Is this procedure correct and results in a sound estimate of the prediction error? Why?
b) How would you change in this procedure? Please write a pseudo-code to elaborate your new procedure. Suppose that the range of λ is given by grid=10^seq (5,- 3,length=50).

a) The procedure is wrong because after the scientist standardize the features using the whole dataset, each standardized feature of any observation will contain the information of the feature values of other observations. The cross-validation in the Step 2 will yield pseudo-training and pseudo-test datasets, which are not non-overlapping. The estimate of the test prediction error will be biased.

(b) (1) Set $Err(\lambda) = 0$, $\lambda \in grid$.
(2) Randomly split the 180 observations into 5 non-overlapping groups with the equal size 36.
(3) For ($i = 1$ to 5) {
(a) The ith group is set as the pseudo-test data set (pTe) and the remaining data as the pseudo-training data set (pTr).
(b) For ($\lambda \in grid$) {

(I) For each feature, compute the sample mean and sample standard deviation with the pTr data and use them to standardize both pTr and pTe. The standardized pseudo-training and pseudo-test datasets are denoted by $\widetilde{pTr}$ and $\widetilde{pTe}$, respectively. Find the LASSO estimate $\hat{\beta}_\lambda$ with λ and the $\widetilde{pTr}$ data.

(II) With the $\hat{\beta}_\lambda$ in (I), compute the test RSS on the $\widetilde{pTe}$ data:

$$test\ RSS = \sum_{i^* \in \widetilde{pTe}} [y^{(i^*)} - \widehat{y^{(i^*)}}]^2,$$

where $\widehat{y^{(i^*)}}$, the prediction for the i^*th test point.

(III) Set

$$Err(\lambda) = Err(\lambda) + test\ RSS.$$

}
}
(4) We choose the best value λ^* with the smallest CV error, i.e. $\lambda^* = \arg\min_\lambda Err(\lambda)$.

3) A classification tree is built for some dataset with the following R code and result:

```
node), split, n, deviance, yval (yprob)
      * denotes terminal node
1) root z1 z2 z3 (z4 z5)
  2) x1 < 633.5  41034  24061.40  0 (0.9139982 0.0860018)
    4) x2 < 0.5  39699  19607.03  0 (0.9325676 0.0674324) *
    5) x2>=0.5   1335   1747.37  1 (0.3617978 0.6382022)
      8) x1 < 162.5  349   431.83  0 (0.6905444 0.3094556) *
      9) x1>=162.5  986  1098.93  1 (0.2454361 0.7545639) *
  3) x1>=z6       2606   3455.89  1 (0.3779739 0.6220261)
    6) x1 < 892.5  1792  2472.00  0 (0.5412946 0.4587054) *
    7) x1>=892.5   814   149.54  1 (0.0184275 0.9815725) *
```

a) What are the missing numbers (z1, . . . , z6) in the R result above? Remark: the deviance reported in output is defined as

$$-2(n_1\, ln\, p_1 \,+\, n_0\, ln\, p_0),$$

where ni, pi are numbers of observations and probability in the node for y = i (i = 1, 0).

b) Compute the (in-sample) false positive rate of this classification tree.

a) Z1 to Z6

$$z_1 = 41034 + 2606 = 43640$$

From node 2, there are 37505 (0's) and 3529 (1's)
From node 3, there are 985 (0's) and 1621 (1's)
Hence, there are 38490 (0's) and 5150 (1's)

$$z_4 = \frac{38490}{43640} = 0.8819$$

$$z_5 = \frac{5150}{43640} = 0.1180$$

$$z_3 = 0$$

$$z_2 = -2 * (38490 * \ln z_4 + 5150 * \ln z_5) = 31677.68$$

$$z_6 = 633.5$$

b) From node 9, misclassify 984 x 0.2454361 = 242 (0's) as 1's
From node 7, misclassify 814 x 0.0184275 = 15 (0's) as 1's

$$FPR = \frac{242 + 15}{38490} = 0.006677059$$