Homework 3

Instructions:

1. Answer all the questions in the homework
2. The number represented in parenthesis, e.g. (2pt) represents two points for the question
3. Follow the sub points carefully
    i) Datasets anywhere in the document appear in **bold** letters
    ii) Variables in the datasets appear *italicized*
4. Please submit only one pdf file for the group in canvas by the due date. Only one member of the group makes the submission on canvas. List the group name and members' names in the submitted file itself
5. For each question provide the code and then its immediate output in the pdf document. Do not submit any R project code files (.R extension files).

---

1. Predicting Software Reselling Profits using Linear Regression

Tayko Software is a software catalog firm that sells games and educational software. It started out as a software manufacturer and then added third-party titles to its offerings. It recently revised its collection of items in a new catalog, which it mailed out to its customers. This mailing yielded 2000 purchases. Based on these data, Tayko wants to devise a linear regression model for predicting the spending amount that a purchasing customer will yield. The file **tayko.csv** contains information on 2000 purchases. The description of the variables is given below

Variable	Description
freq	Number of transactions in the preceding year
last_update	Number of days since last update to customer record
web	1 if customer purchased by web order at least once, 0 otherwise
gender	1 if customer is male, 0 otherwise
address_res	1 if it is a residential address, 0 otherwise
address_us	1 if it is a US address, 0 otherwise
Spending (response)	Amount spending by customer in test mailing (dollars)

a) Partition the data into training (80%) and validation (20%) sets. Run a linear regression on the training data. Report all the accuracy measures for the validation data. (5pt)
(**Instruction**: Set seed to 30)

b) Run LOOCV and K-Fold Cross Validation with K = 10. Report the mean and standard deviation of RMSE measure for both the cross-validation techniques. Compare (smaller or larger etc.) the measures. (20pt)

2. Predicting personal loan acceptance using Logistics Regression

Universal Bank is a relatively young bank growing rapidly in terms of overall customer acquisition. The majority of these customers are liability customers (depositors) with varying sizes of relationship with the bank. The customer base of asset customers (borrowers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business. In particular, it wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors).

A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise smarter campaigns with better target marketing. The goal is to use Logistics Regression to predict whether a new customer will accept a loan offer. This will serve as the basis for the design of a new campaign.

The file **personal_loan.csv** contains data on 5000 customers. The description of the variables is given below

Variable	**Description**
age	Customer age in years
experience	Experience in years
income	Income in thousands of dollars
ccavg	Spending on credit cards
family	Family size
education	Education (undergrad, graduate, advanced)
mortgage	Value of house mortgage in thousands of dollars
securities	1 if customer has securities account with bank, 0 otherwise
cd_account	1 if customer has certificate of deposit account with bank, 0 otherwise
online	1 if customer uses Internet banking facilities, 0 otherwise

credit_card	1 if customer uses credit card issued by the bank, 0 otherwise
personal_loan (response)	accept if customer accepted the loan, reject otherwise

a) Partition the data into training (70%) and validation (30%) sets. Run a logistics regression on the training data. Choose one numeric variable and one binary variable in the model results and interpret their coefficients in odds. (10pt)
(**Instruction**: Set seed to 30)

b) Report the confusion matrix, misclassification rate, overall accuracy, specificity and sensitivity measures for the validation data (10pt)