BUS 41204 Machine Learning

Winter 2023 Midterm

- **This is an INDIVIDUAL exam. You cannot work in groups**.
- The exam must be submitted on gradescope before 11.59pm on Sunday, February 19. This deadline is the same for all sections. You should submit a pdf document.
- Ask coding questions on Ed Discussion Forum. Do not reveal answers when formulating questions.
- When answering questions, provide plots and supporting analysis. Label plots and axes.
- Be concise.
- Provide supporting code to help us understand your answers.

1 Question [30 points]

The file `eBayAuctions.csv` contains information on 1972 auctions transacted on eBay.com during May–June 2004. The goal is to use these data to build a model that will distinguish competitive auctions from noncompetitive ones. A competitive auction is defined as an auction with at least two bids placed on the item being auctioned. The data include variables that describe the item (auction category), the seller (his or her eBay rating), and the auction terms that the seller selected (auction duration, opening price, currency, day of week of auction close). In addition, we have the price at which the auction closed. The goal is to predict whether or not an auction of interest will be competitive.

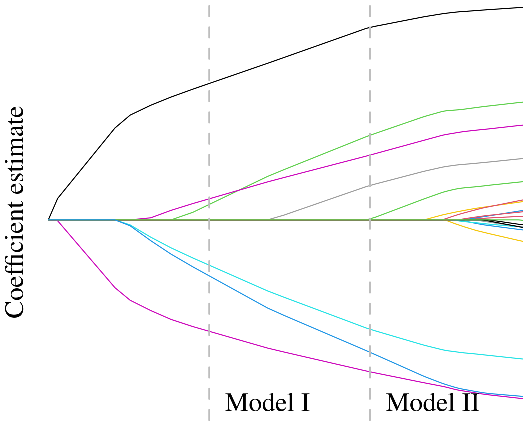
Partition the data into training (60%) and validation (40%) set.

1. Discuss if you can use all the variables to predict at the start of an auction whether it will be competitive.
2. Build a classification tree, a boosted tree, a bagged tree, and a random forest model (with mtry = 4). Choose the tuning parameters for these models by optimizing the performance on the validation set. Report accuracies of these four models as well as their confusion matrices on the validation set.
3. Create lift curves for these four models. What is the lift for the top 10% of observations in the validation set?
4. Briefly describe how boosted tree, bagged tree and random forest models are conceptually similar and how are they conceptually different.

2 Question [32 points]

For each statement write if it is true or false and provide a one sentence explanation. You will get points only if your explanation is correct.

1. The following plot is the solution path for a Lasso estimate. The two dashed lines correspond to two values of the tuning parameter λ, and hence two models.



a. Model I uses a larger λ than Model II.
b. Model I has a higher variance than Model II.
c. Model I has a larger in-sample RSS than Model II.
d. Model I has a larger test error than Model II.

2. Suppose that we would like to fit a linear model to predict MLB players' salaries based on their basic information and statistics for the previous season. The following table shows the residual sum of squares for all the one-predictor models.

Model	**RSS**
$\widehat{\texttt{Salary}} = \beta_0 + \beta_1 \times \texttt{AtBat}$	4.50×10^7
$\widehat{\texttt{Salary}} = \beta_0 + \beta_1 \times \texttt{Hits}$	4.30×10^7
$\widehat{\texttt{Salary}} = \beta_0 + \beta_1 \times \texttt{HmRun}$	4.70×10^7
$\widehat{\texttt{Salary}} = \beta_0 + \beta_1 \times \texttt{Runs}$	4.39×10^7
$\widehat{\texttt{Salary}} = \beta_0 + \beta_1 \times \texttt{RBI}$	4.25×10^7
$\widehat{\texttt{Salary}} = \beta_0 + \beta_1 \times \texttt{Walks}$	4.28×10^7
$\widehat{\texttt{Salary}} = \beta_0 + \beta_1 \times \texttt{Years}$	4.47×10^7
$\widehat{\texttt{Salary}} = \beta_0 + \beta_1 \times \texttt{League}$	5.33×10^7
$\widehat{\texttt{Salary}} = \beta_0 + \beta_1 \times \texttt{Division}$	5.13×10^7
$\widehat{\texttt{Salary}} = \beta_0 + \beta_1 \times \texttt{PutOuts}$	4.85×10^7
$\widehat{\texttt{Salary}} = \beta_0 + \beta_1 \times \texttt{Assists}$	5.32×10^7
$\widehat{\texttt{Salary}} = \beta_0 + \beta_1 \times \texttt{Errors}$	5.33×10^7

Based on the information given by this table, it is certain that

a. The two-variable model identified by forward stepwise selection includes `RBI`.
b. The two-variable model identified by forward stepwise selection includes `Walks`.
c. The two-variable model identified by best subset selection includes `RBI`.
d. The two-variable model identified by backward stepwise selection includes `RBI`.

3. Do the following action increases the bias of the model?

a. Increase k, the number of nearest neighbors, in a kNN classifier.
b. Add interaction terms to a linear regression model.
c. Make more splits when growing a decision tree.
d. Increase B, the number of trees in a random forest model.

4. Suppose that we fit a decision tree to predict a person's height based on gender and age using the following dataset.

Gender	Age	Height (in)
M	20	69
F	21	62
F	15	60
M	19	68

a. The first split is on age.
b. The second split is on age.
c. With two splits, the RSS is smaller than 1.
d. With three splits, the RSS is 0.

3 Question [30 points]

Your company has hired a consulting firm to help with the fraud problem. You are present in the meeting where the consulting firm presents the results of their pilot study, showing that the model has a very low error rate (percent incorrectly classified instances). They argue to your boss that based on this great performance, she should hire them to build the system. You need to explain to your boss the notions of false positive and false negative errors, and how the system should be evaluated. You may assume that the only relevant decision is (binary): if the system predicts fraud, block the account; if the system predicts no fraud, do nothing.

Explain the following notions:

a) describe the confusion matrix to your boss;
b) describe how you will fill out the confusion matrix for the consultant's model;
c) describe the cost/benefit matrix for this problem;
d) explain briefly why the confusion matrix and the cost/benefit matrix are important for this problem (1-2 sentences);
e) show the proper evaluation function (equation) for the consultant's model;
f) how do the confusion and cost matrices come into play in this function.

4 Question [40 points]

You will use data in `marketing.csv` for this question. This dataset contains `64,000` customers who last purchased within twelve months. The customers were involved in an e-mail test: 1/3 were randomly chosen to receive an e-mail campaign featuring a discount offer; 1/3 were randomly chosen to receive an e-mail campaign featuring a buy one get one free offer; 1/3 were randomly chosen to not receive an e-mail campaign. During a period of two weeks following the e-mail campaign, results were tracked. The first 8 columns provide individual-level data and `conversion` column is the label that we will try to predict:

- `recency`: months since last purchase
- `history`: $value of the historical purchases
- `used_discount`: indicates if the customer used a discount before
- `used_bogo`: indicates if the customer used a buy one get one before
- `zip_code`: class of the zip code as Suburban/Urban/Rural
- `is_referral`: indicates if the customer was acquired from referral channel
- `channel`: channels that the customer is using, Phone/Web/Multichannel
- `offer`: the offers sent to the customers, Discount/But One Get One/No Offer

The data were collected through an experiment that allows us to find growth opportunities. Splitting the customers who we are going to send the offer into test (groups receiving one of the two offers) and control groups helps us to calculate incremental gains of offers.

a) Does giving an offer increase conversion? If yes, what kind of offer performs best? Discount or Buy One Get One?

For the rest of the question, we will focus on the customers who received the `Discount` offer or did not receive any offer.

b) Partition the data into 60% train and 40% validation set.

c) Build a model for $P(\text{conversion} = 1 \mid \text{offer}, x)$, where x denotes the individual-level characteristics. Discuss how you chose the parameters used to build the model. Which one is it? Why did you choose it?

Our interest is not just in how different offers did overall, nor is it whether we can predict the probability that a customer will convert after receiving the offer. Rather our goal is to predict how much (positive) impact the offer will have on a specific customer. That way the marketing campaign can direct its limited resources towards the customers who are the most persuadable—those for whom sending the offer will have the greatest positive effect.

d) For each record in the validation set, compute the uplift defined as

$$\text{uplift}(x) = P(\text{conversion} = 1 \mid \text{offer} = \text{Discount}, x) - P(\text{conversion} = 1 \mid \text{offer} = \text{No Offer}, x).$$

If a campaign has the resources to target 25% of the customers, what uplift cutoff should be used? If we are to target 25% of the customers in the validation set based on the uplift obtained from your model, how much better would we do compared to random targeting?