15.071: The Analytics Edge **Fall 2022**
Homework 4: CART and Ensemble Methods

Out: October 19th; Due: November 2nd, 11:59pm.
Please post the assignment in pdf format with file name "Lastname_15071_HW4.pdf".
For each question, please include the main `R` *commands that you used in your submission.*

Problem 1: Preventing Hospital Readmissions [50 pts]

A key performance metric for hospitals is the *30-day unplanned readmission rate*—the proportion of patients discharged from the hospital who had an unplanned readmission within 30 days. Programs like the Hospital Readmissions Reduction Program (HRRP) apply penalties (up to a 3% reduction in payments) to underperforming U.S. hospitals—resulting in withheld payments in excess of $500 million in 2018.

Hospitals can employ some low-cost strategies to reduce unplanned readmissions, such as confirming patient follow-up plans prior to discharge and asking patients to verbally repeat their treatment directions. However, other approaches are more involved and costly. One example is to arrange "telehealth" interventions, in which health care providers contact patients routinely after discharge. Given the cost of these interventions, they are only appropriate for patients at elevated risk of readmission.

You are working for a mid-sized hospital in the northeast United States, and are tasked to assess the impact of telehealth interventions on diabetic patients—with the ultimate goal of reducing the 30-day readmission rate. The intervention will cost approximately $1,200 per patient. Clearly, it must be limited in scope, and a key component of your strategy will be targeting the "right" patients.

Unfortunately, your hospital does not document 30-day readmissions, as this requires significant follow-up with discharged patients. You will thus use a publicly-available dataset to study readmission risk. The dataset includes over 100,000 hospital discharges of over 70,000 diabetic patients from 130 hospitals across the United States during the period 1999–2008.[1] All patients were hospital inpatients for 1–14 days, and received both lab tests and medications while in the hospital. The 130 hospitals represented in the dataset vary in size and location: 58 are in the northeast United States and 78 are mid-sized (100–499 beds).

The dataset is provided in the `readmission.csv` file. It contains the following variables:

- **readmission**: 1 if the patient had an unplanned readmission within 30 days, 0 otherwise.
- **Patient characteristics**: `race`, `gender`, and `age` capture demographic information.
- **Recent medical system use**: The variables `numberOutpatient`, `numberEmergency`, and `numberInpatient` capture the number of times the patient used the medical system in the last year.
- **Diabetic treatments**: A number of variables capture the patient's diabetic treatments: `acarbose`, `chlorpropamide`, `glimepiride`, `glipizide`, `glyburide`, `glyburide.metformin`, `insulin`, `metformin`, `nateglinide`, `pioglitazone`, `repaglinide`, and `rosiglitazone`.
- **Admission information**: The variables `admissionType` and `admissionSource` contain information about how the patient was admitted to the hospital. The variable `numberDiagnoses` captures the number of diagnoses the patient had recorded for their admission. There are also a number of variables that indicate whether a patient was diagnosed with various conditions when admitted: `diagAcuteKidneyFailure`, `diagAnemia`, `diagAsthma`, `diagAthlerosclerosis`, `diagBronchitis`, `diagCardiacDysrhythmia`, `diagCardiomyopathy`, `diagCellulitis`, `diagCKD`, `diagCOPD`, `diagDyspnea`, `diagHeartFailure`, `diagHypertension`, `diagHypertensiveCKD`,`diagIschemicHeartDisease`, `diagMyocardialInfarction`, `diagOsteoarthritis`, `diagPneumonia`, and `diagSkinUlcer`.

[1] Beata Strack, Jonathan P. DeShazo, Chris Gennings, et al., "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," *BioMed Research International*, vol. 2014, 2014.

- **Treatment information**: `timeInHospital` is the number of days the patient was in the hospital, and `numLabProcedures`, `numNonLabProcedures`, and `numMedications` capture the amount of care the patient received in the hospital.

Run the following command to read in the dataset

```
readmission = read.csv("readmission.csv")
set.seed(144)
split = createDataPartition(readmission$readmission, p = 0.75, list = FALSE)
readm.train <- readmission[split,]
readm.test <- readmission[-split,]
```

a) Perform some exploratory data analysis on the training data set and report two interesting insights you gained from your analysis. [**4 pts**]

b) Based on conversations with the hospital's management, you estimate the cost of a 30-day unplanned readmission at $35,000. From published information at a similar institution, you estimate that telehealth interventions will reduce the incidence of 30-day unplanned readmissions in the treated population by 25%. Given the cost of $1,200 per intervention, what are:

- – the difference in costs between a false positive and a true negative;
- – -the difference in costs between a false negative and a true positive.

Define the loss matrix for your CART model based upon these differences in cost [**10 pts**]

c) Fit a CART model using a `cp` parameter of 0.001 and the loss matrix defined in Question b. Include an image of your tree. [**10 pts**]

d) Assess the model's predictive performance using the test set. What is the accuracy, true positive rate and false positive rate? Write a short memo contrasting the decisions resulting from your model and those resulting from current practice (under which no patient is subject to a telehealth intervention). Provide summary statistics to explain how the decisions differ, and discuss the costs and benefits of each approach. Make sure to compare the total monetary costs of patient readmission. [**10 pts**]

e) Since your hospital has yet to pilot a telehealth intervention, your analysis has relied on external estimates of the cost of the intervention ($1,200 per patient). However, this estimate may be somewhat inaccurate for your hospital. Using the classifications returned by the CART tree in part c), investigate how the expected cost reductions would change as the cost of the intervention varies. For which range of values for the true cost of the intervention would the CART classifier from c) return cost-effective recommendations? [**10 pts**]

f) Coming up with a strong classifier to predict 30 day hospital readmissions could be of interest to the hospital for many reasons beyond the intervention described in this problem. Suppose the hospital instead wanted you to create a CART tree with strong overall discriminative ability, that could be used for a host of applications with potentially different losses for false positives and false negatives. How might we train a CART tree with the objective of strong overall performance in mind, rather than creating a tree that does well for a particular loss function? [**6 pts**]

Problem 2: Predicting Housing Prices in Ames, Iowa [50 pts]

In this problem, you will compare three predictive analytics methods: linear regression, CART, and random forest.

Note: Please use the **ames.csv** file for this problem.

First, import the dataset and split it into a training set (70%) and a test set (30%) using the following commands.

```
ames = read.csv("ames.csv")
set.seed(15071)
split = createDataPartition(ames$SalePrice, p = 0.7, list = FALSE)
ames.train = ames[split,]
ames.test = ames[-split,]
```

a. Construct a linear regression model with all the independent variables. Also, notice that some coefficients are reported as “NA”. Why is this happening? [**5 pts**]

b. Create a CART tree. Use cross-validation to select the value of the **cp** parameter. What is the selected value of **cp**? Provide an image of your tree. Comment on the role of the different variables used in the model. [**10 pts**]

c. Construct a random forest model with 80 trees and a **nodesize** of 25. Use cross-validation to select the value of the **mtry** parameter. What is the selected value of **mtry**? Which variables are most important in the model? Please comment briefly. [**10 pts**]

d. For each of the three models, report 6 performance metrics: the in-sample R^2, MAE and $RMSE$ and the out-of-sample R^2, MAE and $RMSE$. [**10 pts**]

e. Suppose instead of tuning **mtry** in the random forest algorithm using cross-validation, we instead tuned **mtry** based upon optimal performance on the test set. Suppose we then compared the out of sample R^2 between random forests, the linear regression fit in part a), and the CART tree fit in part b). Would this be a fair reflection of the differences in the quality of predictions furnished by these three algorithms for future, unseen data? Explain. [**5 pts**]

f. Which of the three models would you recommend? Make sure to justify your choice and to discuss the strengths and limitations of the three models. [**10 pts**]

Problem 3: eBay.com [50 pts]

The file **`eBayAuctions.xls`** 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 classify competitive auctions from noncompetitive ones. A competitive auction is defined as an auction with at least two bids placed on the item auctioned. The data include variables that describe the item (auction category), the seller (his/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 the auction will be competitive.

Data Preprocessing. Create dummy variables for the categorical predictors. These include Category (18 categories), Currency (USD, GBP, Euro), EndDay (MondaySunday), and Duration (1, 3, 5, 7, or 10 days). Split the data into training and validation datasets using a 60% : 40% ratio. Make sure you are doing classification and note that there is a slight class imbalance you need to account for.

(a) Run AdaBoost (R function: **`adabag::boosting`**), XGBoost (R function: **`xgboost::xgboost`**), bagging (R function: **`xgboost::bagging`**), random forest (R function: **`randomForest::randomForest`**) **with all predictors**. Fill in the table below with training accuracy, validation accuracy, and the most important variable (R function: **`adabag::importanceplot`**, **`xgboost::xgb.importance`**, and **`randomForest::varImp`**). **[12 pts]**

Table 1: Results with all predictors

	AdaBoost	XGBoost	bagging	random forest
Training accuracy				
Validation accuracy				
Most important variable				

(b) Run AdaBoost, XGBoost, bagging, random forest **with all predictors except "ClosePrice"**. Fill in the table below with training accuracy, validation accuracy, and the most important variable. **[10 pts]**

Table 2: Results with all predictors except "ClosePrice"

	AdaBoost	XGBoost	bagging	random forest
Training accuracy				
Validation accuracy				
Most important variable				

(c) Run AdaBoost, XGBoost, bagging, random forest **with all predictors except "OpenPrice"**. Fill in the table below with training accuracy, validation accuracy, and the most important variable. **[10 pts]**

Table 3: Results with all predictors except "OpenPrice"

	AdaBoost	XGBoost	bagging	random forest
Training accuracy				
Validation accuracy				
Most important variable				

(d) Run AdaBoost, XGBoost, bagging, random forest **with all predictors except "OpenPrice" and "ClosePrice"**. Fill in the table below with training accuracy, validation accuracy, and the most important variable. [**10 pts**]

Table 4: Results with all predictors except "OpenPrice" and "ClosePrice"

	AdaBoost	XGBoost	bagging	random forest
Training accuracy				
Validation accuracy				
Most important variable				

(e) Compute the proportion of the majority class in the test set. Comment on the result from (a), (b), (c), and (d). Which one you should use in reality and why? [**10 pts**]