HW3 – Solution/Rubric

Question #1

(a)(10 points)

$$\Delta = \mathrm{C}_{\mathrm{imp}}(T_{old}) - C_{imp}(T_{new})$$

$$= \sum_{m=1}^{M} N_m Q_m(T_{old}) - \sum_{m=1}^{M+1} \widetilde{N_m}\tilde{Q}_m(T_{new}) \qquad (1)$$

$$= \sum_{\mathrm{m=1}}^{\mathrm{M}} \sum_{i:x_i\in R_m} (y_i - \widehat{y_m})^2 - \sum_{\mathrm{m=1}}^{\mathrm{M+1}} \sum_{i:x_i\in \widetilde{R_m}} \left(y_i - \widehat{\widetilde{y_m}}\right)^2 \qquad (2)$$

Since the first M-1 regions are the same in the new tree and the old tree,

$$= \sum_{i:x_i\in R_M} (y_i - \widehat{y_M})^2 - \sum_{\mathrm{m=M}}^{\mathrm{M+1}} \sum_{i:x_i\in \widetilde{R_m}} \left(y_i - \widehat{\widetilde{y_m}}\right)^2 \qquad (3)$$

Note that $\mathrm{R_M} = \widetilde{R_M} \cup \widetilde{R_{M+1}}$, and

$$\widehat{y_M} = \frac{1}{N_M} \sum_{i:x_i\in \mathrm{R_M}} y_i \qquad (4)$$

$$\widehat{\widetilde{y_m}} = \frac{1}{\tilde{N}_m} \sum_{i:x_i\in \tilde{R}_{\mathrm{m}}} y_i \qquad (5)$$

(3) only depend on the data points that fall in region $\mathrm{R_M}$ in the old tree.

Grading criteria:

- **2 points**: For getting (1)
- **2.5 points:** For getting (2)
- **5 points:** For getting (3)
- **0.5 point:** For explaining why (3) only depend on data points that fall in region $\mathrm{R_M}$. Or further to a formula that is similar to (3) but with (4) and (5) plugged in.
- Full credits if provide (3) or that formula directly without providing (1) or (2).

(b)(10 points)

Define the function $\mathrm{RSS}_{\{\mathrm{y_i}\}_{i\in\{1\ldots n\}}}(z) = \sum_{i=1}^{n}(z_i - z)^2$

$$\Delta = \sum_{i:x_i\in R_M} (y_i - \widehat{y_M})^2 - \sum_{\mathrm{m=M}}^{\mathrm{M+1}} \sum_{i:x_i\in \widetilde{R_m}} \left(y_i - \widehat{\widetilde{y_m}}\right)^2$$

$$= \sum_{i:x_i\in R_M} (y_i - \widehat{y_M})^2 - \sum_{i:x_i\in \widetilde{R_M}} \left(y_i - \widehat{\widetilde{y_M}}\right)^2 - \sum_{i:x_i\in \widetilde{R_{M+1}}} \left(y_i - \widehat{\widetilde{y_{M+1}}}\right)^2$$

Since $\mathrm{R_M} = \widetilde{R_M} \cup \widetilde{R_{M+1}}$,

$$= \sum_{i:x_i\in \tilde{R}_M} (y_i - \widehat{y_M})^2 + \sum_{i:x_i\in \tilde{R}_{M+1}} (y_i - \widehat{y_M})^2 - \sum_{i:x_i\in \widetilde{R_M}} \left(y_i - \widehat{\widetilde{y_M}}\right)^2 - \sum_{i:x_i\in \widetilde{R_{M+1}}} \left(y_i - \widehat{\widetilde{y_{M+1}}}\right)^2$$

$$= \sum_{i:x_i \in \tilde{R}_M} (y_i - \widehat{y_M})^2 - \sum_{i:x_i \in \widetilde{R_M}} \left(y_i - \widetilde{\widehat{y_M}}\right)^2 + \sum_{i:x_i \in \tilde{R}_{M+1}} (y_i - \widehat{y_M})^2 - \sum_{i:x_i \in \widetilde{R_{M+1}}} \left(y_i - \widetilde{\widehat{y_{M+1}}}\right)^2$$

$$= \mathrm{RSS}_{\{\mathrm{y_i}\}_{\mathrm{i:x_i} \in \tilde{R}_\mathrm{M}}} (\widehat{y_M}) - \mathrm{RSS}_{\{\mathrm{y_i}\}_{\mathrm{i:x_i} \in \tilde{R}_\mathrm{M}}} \left(\widetilde{\widehat{y_M}}\right) + \mathrm{RSS}_{\{\mathrm{y_i}\}_{\mathrm{i:x_i} \in \tilde{R}_\mathrm{M+1}}} (\widehat{y_{M+1}}) - \mathrm{RSS}_{\{\mathrm{y_i}\}_{\mathrm{i:x_i} \in \tilde{R}_\mathrm{M+1}}} \left(\widetilde{\widehat{y_{M+1}}}\right) \quad (6)$$

Note that $\widetilde{\widehat{y_M}} = \frac{1}{\tilde{N}_M} \sum_{i:x_i \in \tilde{R}_\mathrm{M}} y_i$ and $\widetilde{\widehat{y_{M+1}}} = \frac{1}{\tilde{N}_{M+1}} \sum_{i:x_i \in \tilde{R}_\mathrm{M+1}} y_i$, from the hint we know that they are the minimizer of $\mathrm{RSS}_{\{\mathrm{y_i}\}_{\mathrm{i:x_i} \in \tilde{R}_\mathrm{M}}}(z)$ and $\mathrm{RSS}_{\{\mathrm{y_i}\}_{\mathrm{i:x_i} \in \tilde{R}_\mathrm{M+1}}}(z)$, respectively.

Hence $\mathrm{RSS}_{\{\mathrm{y_i}\}_{\mathrm{i:x_i} \in \tilde{R}_\mathrm{M}}} (\widehat{y_M}) - \mathrm{RSS}_{\{\mathrm{y_i}\}_{\mathrm{i:x_i} \in \tilde{R}_\mathrm{M}}} \left(\widetilde{\widehat{y_M}}\right) \geq 0$ (7)

And $\mathrm{RSS}_{\{\mathrm{y_i}\}_{\mathrm{i:x_i} \in \tilde{R}_\mathrm{M+1}}} (\widehat{y_{M+1}}) - \mathrm{RSS}_{\{\mathrm{y_i}\}_{\mathrm{i:x_i} \in \tilde{R}_\mathrm{M+1}}} \left(\widetilde{\widehat{y_{M+1}}}\right) \geq 0$ (8)

So we have $\Delta \geq 0$.

Grading criteria:

- **5 points**: For getting (6).
- **3 points:** For saying that $\widetilde{\widehat{y_M}}$ and $\widetilde{\widehat{y_{M+1}}}$ are the minimizers.
- **2 points:** For getting (7) and (8).

(c)(10 points)

(=>)

$$\mathrm{C}_\alpha(T_{new}) - C_{\alpha(T_{old})} \leq 0$$

$$\Leftrightarrow \mathrm{C_{imp}}(\mathrm{T_{new}}) - \mathrm{C}_{imp}(\mathrm{T_{old}}) + \alpha\mathrm{SST} \leq 0 \qquad (9)$$

$$\Leftrightarrow -\Delta + \alpha\mathrm{SST} \leq 0 \qquad (10)$$

plug in (2)

$$\Leftrightarrow \sum_{\mathrm{m=1}}^{\mathrm{M}} \sum_{i:x_i \in R_m} (y_i - \widehat{y_m})^2 - \sum_{\mathrm{m=1}}^{\mathrm{M+1}} \sum_{i:x_i \in \widetilde{R_m}} \left(y_i - \widetilde{\widehat{y_m}}\right)^2 \geq \alpha SST \quad (11)$$

by the definition of SSE,

$$\Leftrightarrow \mathrm{SSE_{old}} - \mathrm{SSE_{new}} \geq \alpha SST \qquad (12)$$

$$\Leftrightarrow \frac{\mathrm{SSE_{old}}}{SST} - \frac{\mathrm{SSE_{new}}}{SST} \geq \alpha$$

$$\Leftrightarrow \mathrm{R^2_{new}} - \mathrm{R^2_{old}} \geq \alpha$$

Grading criteria:

- **3 points**: For getting (10)
- **2 points:** For getting (12)

- **5 points:** For getting the final result.

Question #2 (70 points)

Remark: all the answers in a ballpark range should be allowed. R 3.5 and 3.6 with the same seed will result in different train/test split.

(25 points) This exercise is focused on determining whether a letter is B or not.

i) **(3 points)** Baseline model is to predict “not B”. The accuracy of the baseline is:

$$Accuracy = \frac{818}{818 + 273} \approx 74.98\%$$

Grading criteria:

- **3 points**: For getting accuracy correct (or in ballpark range).

ii) **(5 points)** For logistic regression model, we don’t explicitly ask to show the model. However, if it is included, that can be helpful (especially if they do not get the same results as we do).

When running their model and afterwards calculating the accuracy, a value of ≈ 93.6% was reached.

Grading criteria:

- **5 points**: For getting accuracy correct (or in ballpark range $\pm 0.5\%$).

NOTE: If result is incorrect, but intermediate calculations (or logistic model) are included, award up to 2 points. Otherwise, assign 0 points.

iii) **(2 points)** We calculate the area under the curve (AUC)recovering a value of 0.978.

Grading criteria:

- **2 points**: For getting AUC correct (or in ballpark range ± 0.03).

iv) (5 points) To build the CART model, we found the c_p value using 5-fold cross validation. This implies trying an array of possible values for c_p (0:0.001:0.1) and, for each value of c_p, we estimated the OSR^2 by training 5 different CART models by leaving 1/5 of the training data out and testing the model performance on the left out data. We averaged the 5 estimates for each c_p and kept the value of c_p that maximized the model accuracy. In our case, the best value was 0.004.

Once the value of c_p was calculated, we then trained the CART model for that value using all of the training data to build our final model. When applying this model on the test data, we reached an accuracy of $\approx$ 93.9%.

Grading criteria:

- **2 points**: For explanation on how cross-validation was carried out and how c_p was selected.
- **1 point:** For getting $c_p = 0.004$ value correct (Different seed may lead to slightly different result, deduct points only when the result is abnormal).
- **2 points:** For getting accuracy correct (Different seed may lead to slightly different result, deduct points only when the result is abnormal).

v) (5 points) Training our Random Forest model, just like we did in class, using the default parameter values, we get an accuracy of $\approx$ 97.6%.

Grading criteria:

- **3 points**: For getting accuracy correct (Full credits for reaching 96%).
- **2 points**: For using default parameters.

vi) (5 points) Summarizing our findings for the models in part (b):

Model	**Accuracy**
Logistic Regression	93.6%
CART	93.9%
Random Forest	97.6%

Given these results, random forest performs best on the test set. For this application, given that we are processing sensitive information (checks at ATMs, for example), we are interested in maximizing accuracy instead of interpretability.

Grading criteria:

- **1 points**: For correctly concluding that random forest performs better

- **2 points:** For correctly mentioning that we should focus on accuracy
- **2 points**: For explanation that relates back to application! (discount 1 point if it doesn't refer to applications).

- **(45 points)** This exercise is now relating to multiclass classification (A, B, P, R)

(i) **(5 points)** Baseline model is to predict "A" based on training data:

A	B	P	R
525	493	520	487

When applying our baseline model on the test data, we get an accuracy of:

$$Accuracy = \frac{264}{264+273+283+269} \approx 24.2\%$$

Grading criteria:

- **1 points**: For using training data for building baseline.
- **2 points:** For correctly stating baseline is to predict "P"
- **2 points:** For correctly calculating baseline accuracy (or in ballpark range ±0.5%).

(ii) **(5 points)** The test set accuracy for the trained LDA model is ≈ 91.9%.

```
pred_test_lda_class

     A   B   P   R
A  249   2   4   9
B    1 235   0  37
P    0  10 272   1
R    0  24   0 247
```

(iii) **(5 points)** To build the CART model, we found the c_p value using 5-fold cross validation in the same manner as part (b-iv). The optimal c_p found was 0 (searched over grid (0:0.001:0.1)) based on maximizing accuracy. The test set accuracy for the trained model is ≈ 91.1%. In general, there is no guarantee that they would get generally the exact same results for most things (including hyperparameter tuning) due to the randomization. All reasonable answers should get full credit.

Grading criteria:

- **2 points**: For explanation on how cross-validation was carried out and how c_p was selected (it is ok to reference part (b-iv))
- **1 point:** For getting correct c_p value. (Different seed and cp grid leads to different result, but cp should not be very large) Points only deducted for abnormal result.
- **2 points:** For getting accuracy correct.

iv)(5 points) Training our Random Forest model, setting mtry = 16, we get an accuracy of ≈ 96.52%.

Grading criteria:

- **2 points**: For getting accuracy correct.
- **3 points**: For using the correct mtry value.

v) (8 points) To build the new RF model, we found the m_{try} value using 5-fold cross validation. This implies trying an array of possible values for m_{try} (1:16) and, for each value of m_{try}, we estimated the OSR^2 by training 5 different RF models by leaving 1/5 of the training data out and testing the model performance on the left out data. We averaged the 5 estimates for each m_{try} and kept the value of m_{try} that maximized the model accuracy. In our case, the best value was 2.

Once the value of m_{try} was calculated, we then trained the RF model for that value using all of the training data to build our final model. When applying this model on the test data, we reached an accuracy of ≈ 98.0%.

Grading Criteria:

- **3 points**: For explanation on how cross-validation was carried out and how m_{try} was selected (it is not ok to reference part (b-iv), because there were changes now).
- **3 point:** For getting $m_{try} = 2$ value correct (or in ballpark range ± 1).
- **2 points:** For getting accuracy correct (or in ballpark range $\pm 0.1\%$).

vi)(7 points) Based on the hints provided in the problem statement, we are able to run boosting in a straightforward manner. The test accuracy found is ≈ 98.17%.

Grading criteria:

- **7 points:** for correct accuracy

NOTE: Provide partial credit (up to 3 points) if result is incorrect but intermediate steps provided (such as confusion matrix, for example; basically anything that allows us to better understand where they might have gone wrong)

vii)(10 points) Summarizing our findings for the models in part (c):

Model	**Accuracy**
LDA	91.9%
CART	91.1%
Bagging	96.52%
Random Forest	98.0%
Boosting	98.17%

The application for multiclass classification hasn't changed, since we are interested in being as accurate as possible.

Below two answers are both correct:

1. together with the fact that RF can be parallelized, means that we will not change our selected model compared to part (a): use random forest!
2. Since boosting achieves the highest accuracy, use boosting.

Grading criteria:

- **4 points**: for correctly identifying the best model.
- **3 points**: for correctly identifying that accuracy is most important metric.
- **3 points**: for explanation that relates back to applications at hand
  NOTE: For the last two bullets, it suffices to refer to last question in part (a).