DELIVERABLE 2

Due by Monday, September 29th, 23:59

This is a team deliverable. Work with your assigned team members only. Each team needs to upload its solutions to Canvas as a single PDF file (any team member can do so through their Canvas account). All team members should contribute to the deliverable. If a team member did not contribute, then their name should not appear on the cover page. Scripts should be written in ***R****.*

Problem 1 (25 points) • Decision Trees for Regression

For this problem we are interested in predicting healthcare charges, in USD, using a patient's age and BMI as features. Please load the **`insurance_charges.csv`** file provided. [You do not need to do a training/test split for this problem].

a) Fit a decision tree to predict **`charges`** using only **`bmi`** and **`age`** as independent variables. Leave all the decision tree's hyperparameters at their default values. Report the model's R^2 score and plot the decision tree.

b) Fit a decision tree to predict **`charges`** using only the features **`f_bmi`** and **`age`**. The feature **`f_bmi`** is related to the feature **`bmi`**. Again, leave all hyperparameters of the decision tree at their default values and report the R^2 score of the model and plot the decision tree.

c) Comment on how the decision trees in parts (a) and (b) compare. Make sure to compare both the performances of the trees and their structures. Please explain why this phenomenon occurs. Take your best guess on what is the relationship between feature **`f_bmi`** and feature **`bmi`**. Hint: Plot these two features.

d) Can you explain mathematically why you observe the phenomenon in part (c)?

e) Now, we are interested in predicting a new feature **`cardiovascular_care_cost`**, the amount in USD a patient spends on cardiovascular care, using a CART tree. Take a range of values for the **`cp`** parameter {0, 0.02, 0.04, 0.06,0.08, 0.1}, fit a CART tree using each **`cp`** parameter, and comment on how well they fit. Provide an explanation about how the model fit changes with **`cp`**.

   [Hint: Consider ways to visualize how the tree fits change as **`cp`** changes].

Problem 2 (25 points) • The Power of "Wise" Randomness

We'll continue with the dataset of Problem 1.

a) Split **insurance_charges.csv** into a 70% train and 30% test dataset. Use the **glimpse** function in **R** to both the training and test datasets, and show your output. [Note: For reproducible results, use **set.seed**(15072) to set the seed of the random number generator before splitting into train/test.]

b) Fit a decision tree of depth no greater than 4 to the training dataset (from part (a)) using all available features to predict charges. Do not change the other hyperparameters in the function used to fit the tree. Plot the tree and show your result. Additionally, report the R^2 of this tree, called the **basetree**, on the test dataset.

c) Consider the residuals between the charges predicted by the tree from part b) and the actual charges on the test dataset. Apply the **summarize** function on these residuals, and report the output. Plot a histogram of the residuals. [Note: In part (g), you will be comparing these results with the results from part (f).]

d) Take a simple random sample with replacement of size 50 from the training dataset (i.e. in a sample of 50 datapoints, there may be repeats). You can use the **sample()** function in **R** for this. Fit a decision tree with depth no greater than 4 to this sample. Plot the tree. Report the $\mathbf{R}^2$ of this tree on the test set.

e) Perform the process in part d) for a total of 30 times. Note that these trees will be different due to the sampling process. Plot a histogram of the 30 R^2 values on the test set. Comment on how the test performance of these 30 trees compare to that of the **basetree**.

f) Consider **WiseTree**: For an input **x** the prediction of **WiseTree** is given by the average of the predictions at **x** from each of the 30 trees. Use the **summarize** function on the residuals formed between **WiseTree** predictions and actual charges (on the test set). Report the output. Plot a histogram of these residuals.

g) Report the R^2 of **WiseTree** on the test dataset. Using the results from parts c) and f) comment on which model (**basetree** or **WiseTree**) has a better prediction performance on the test set. Provide an explanation in 2-3 sentences.

Problem 3 (25 points) • Logistic Regression Redux

In this problem, we consider a business application: measuring customer churn for Watson Analytics. The dataset for this question describes usage behavior of 7,032 users across 1 year.

You will work with the dataset provided in **`customerchurn.csv`**. It contains 7 variables:

- **`Churn`**: the user's usage status at year-end (i.e. 1 is churn and 0 is not churn)
- **`MonthlyCharges`**: monetary cost of user's plan
- **`SeniorCitizen`**: whether user age is above 60 years
- **`PaymentMethod`**: channel of payment
- **`InternetService`**: type of internet connection
- **`tenure`**: number of years passed as a user
- **`Contract`**: payment installment terms

a) Use the **`table`** function on the variable **`Churn`** to study some summary characteristics of this variable. How many customers churned? How many customers did not churn? Compute the customers' churn rate.

b) Split the dataset into a 70% train and 30% test set. [Use **`set.seed`**(15072) to set the seed of the random number generator before splitting into train/test]. Train a logistic regression model to predict Churn using all independent variables except **`PaymentMethod`**. State the value of and interpret the coefficient for **`SeniorCitizen`**.

c) Using your model in part b), compute the churn probability for the 5th user in the full churn dataset **`customerchurn.csv`**.

d) Using the model from part b), predict churn on the entire test set. Your manager asks you to set the cutoff probability as 0.3 (any probability higher than 30% will be considered as churn). Calculate and attach the confusion matrix for the test set predictions.

e) What is the meaning of a false positive in this context? false negative?

f) Suppose you are a business analyst at Watson Analytics, and you are faced with a tradeoff. Should you focus on minimizing false negatives or false positives that arise from the churn model? Describe and defend your decision by interpreting the business implications of both false negatives and false positives.

g) Your manager now changes her mind and decides to increase the cutoff probability. How would the false positive rate (FPR) and false negative rate (FNR) change as you change the cutoff probability? Accompany your answer with a plot of how the FPR/FNR changes on this dataset.

h) Your manager estimates the profit of predicted churn status versus actual churn status in the matrix shown below. Given this information, how would you optimize the probability threshold (Note: Consider twenty different threshold values equally spaced between 0 and 1.)

	Actual Churn	Actual Non-Churn
Predicted Churn	$2000	-$1000
Predicted Non-Churn	-$6000	$3000

Problem 4 (25 points) • LinR vs CART vs Random Forest

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

Consider the Ames, Iowa Housing Prices dataset, which describes sales of 2,838 properties in the town of Ames, Iowa from 2006 to 2010[1]. Work with the dataset provided in the **`ames.csv`** file. This file has been pre-processed to simplify the analysis. The file contains 73 variables, described on the last page of this handout. The first variable characterizes the property's sale price, which we aim to predict. The other variables describe the property in detail, both in quantitative terms (square footage, size of the lot, number of rooms, date of construction, etc.) and qualitative terms (building materials, style of the home, etc.).

First, import the dataset and create a 70-30 train-test split. Use 15072 as the random seed throughout your code.

a) Construct a linear regression model on the training set. Given the number of variables in the model and the mix of numerical and categorical variables we recommend using the **`ols_step_backward_p`** function of the package "**`olsrr`**" (discussed in Recitation 2):

```
mod_linear_intial <- lm(SalePrice ~ ., data = df_train)
summary(mod_linear_intial)
mod_linear_olsrr <- ols_step_backward_p(mod_linear_intial, p_val =
0.05, progress = TRUE)
mod_linear_final <- mod_linear_olsrr$model
summary(mod_linear_final)
```

This will take a few minutes but is faster than manual backward elimination. Assume that multi-collinearity is not an issue. What is the in-sample and out-of-sample R-squared for this model?

b) Train a CART model without changing any default parameters. Include an image of your tree in your solution. Which variables seem important? What is the in-sample and out-of-sample R-squared for this model?

c) A friend of yours wants to sell her home in Ames, Iowa and has called you for advice. The house does not have a central air conditioning system. She is wondering whether it would be worth to have one installed in order to increase the value of her home (at a cost of $15,000). What would your recommendation be and what does this say about your models in part (a) and (b)?

d) Use 10-fold cross-validation to select a good cp value for your tree and define your final model. Consider the following cp values: 5×10^{-6}, 5×10^{-5}, 5×10^{-4}, 5×10^{-3}, 5×10^{-2}. Report the value of cp that you select. What is the in-sample and out-of-sample R-squared for the selected model?

e) 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`**? What is the in-sample and out-of-sample R-squared for the selected model?

f) Which of the four models you constructed in parts (a), (b), (d), and (e) would you recommend? Make sure to justify your choice and to discuss the strengths and limitations of the models.

[1] De Cock, Dean. "Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project." *Journal of Statistics Education*, 19.3 (2011)

- SalePrice - the property's sale price (dollars)
- MSZoning: The general zoning classification
- LotFrontage: Linear connected street line (feet)
- LotArea: Lot size (sq. feet)
- Street: Type of road access
- Alley: Type of alley access
- LotShape: General shape of property
- LandContour: Flatness of the property
- LotConfig: Lot configuration
- LandSlope: Slope of property
- Neighborhood: Physical locations within Ames city
- Condition1: Proximity to main road or railroad
- Condition2: Other condition, if applicable
- BldgType: Type of dwelling
- HouseStyle: Style of dwelling
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- RoofStyle: Type of roof
- RoofMatl: Roof material
- Exterior1st: Exterior covering on house
- Exterior2nd: Exterior covering on house (if other)
- MasVnrType: Masonry veneer type
- MasVnrArea: Masonry veneer area (sq. feet)
- ExterQual: Exterior material quality
- ExterCond: Condition of the exterior material
- Foundation: Type of foundation
- BsmtQual: Height of the basement
- BsmtCond: General condition of the basement
- BsmtExposure: Walkout or garden level basement
- BsmtFinType1: Quality of 1^{st} basement area
- BsmtFinSF1: 1^{st} basement finished area (sq. feet)
- BsmtFinType2: Quality of 2^{nd} bsmt. area (if any)
- BsmtFinSF2: 2nd basement finished area (sq. feet)
- BsmtUnfSF: Unfinished basement area (sq. feet)
- TotalBsmtSF: Total basement area (sq. feet)
- Heating: Type of heating
- HeatingQC: Heating quality and conditionCentralAir: Central air conditioning
- Electrical: Electrical system
- X1stFlrSF: First floor area (sq. feet)
- X2ndFlrSF: Second floor area (sq. feet)
- LowQualFinSF: Low quality finished area (sq. feet)
- GrLivArea: Ground living area (sq. feet)
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- BedroomAbvGr: # bedrooms above ground
- KitchenAbvGr: # kitchens above ground
- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Number of rooms above grade
- Functional: Home functionality rating
- Fireplaces: Number of fireplaces
- FireplaceQu: Fireplace quality
- GarageType: Garage location
- GarageYrBlt: Year garage was built
- GarageFinish: Interior finish of the garage
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage (sq. feet)
- GarageQual: Garage quality
- PavedDrive: Paved driveway
- WoodDeckSF: Wood deck area (sq. feet)
- OpenPorchSF: Open porch area (sq. feet)
- EnclosedPorch: Enclosed porch area (sq. feet)
- X3SsnPorch: Three season porch area (sq. feet)
- ScreenPorch: Screen porch area (sq. feet)
- PoolArea: Pool area (sq. feet)
- PoolQC: Pool quality
- Fence: Fence quality
- MiscFeature: Miscellaneous feature
- MoSold: Month Sold
- YrSold: Year Sold
- SaleType: Type of sale