Course 2 Task 2 Project Instructions

You have been asked by Danielle Sherman, CTO of Blackwell Electronics, to predict the customers' brand preferences that are missing from the incomplete surveys by conducting two classification methods in R. Once you have determined which classifier—*k*-nearest-neighbor or decision tree—works better on the provided data set, she would like you to predict the brand preferences for the incomplete survey responses and prepare a report of your findings.

**This task requires you to prepare one deliverable for Danielle Sherman:**

* ***Customer Brand Preferences Report.*** A report in a Zip file that includes:
  + A brief summary in Word or Excel of your methods and results that includes:
  + The classifiers you tried.
  + The classifier you selected to make the predictions, including a rationale for selecting the method you did and the level of confidence in the predictions.
  + The predicted answers to the brand preference question for the instances of survey results that are missing that answer.
  + A chart that displays the customer preference for each brand based on the combination of the actual answers and the predicted answers to the brand preference survey question.
  + The results of each classifier you ran exported from R

#### [2. Introduction to the caret Package](https://ut.daacertificate.com/mc/poa?productID=2655&taskID=3351" \l "collapsepoa4320)

The [**caret**](http://cran.r-project.org/web/packages/caret/index.html) package (short for Classification And REgression Training) is a set of functions that attempt to streamline the process for creating predictive models. The package contains tools for:

* data splitting
* pre-processing
* feature selection
* model tuning using resampling
* variable importance estimation

In this step you will work through the following exercises to become more familiar with this all-important package for R; you will be using caret throughout the remainder of the program so it is very important that you become very confident with applying it to data analytics problems. To do this you will need to work through a short tutorial on the package and be prepared to discuss your questions and findings in the weekly mentor meetings.

To begin:

1. Work through the [A Short Introduction to the caret Package](https://cran.r-project.org/web/packages/caret/vignettes/caret.pdf) tutorial so you can understand how the training process in caret works. You’ll need the skills from the tutorial to work through the remainder of this task so be sure to allocate enough time for working the examples and working through some of the free datasets from the optional resources in task one to reinforce your learning.
2. After working through the introduction to the caret package: Develop a working ‘pipeline’ (an example can be seen in the resources) for model training/testing using the caret package so you can easily train/test different models for the capstone. At a minimum you should utilize the following functions with a definitive training and testing set that you will build from the training data that you have already sampled:
   * createDataPartition()
   * trainControl()
   * train()
   * predict()
   * postResample()
3. You might also find it quite useful to use an alternate tuning grid such as expand.grid, which can be used to specific any of the numerous training parameters that are available with many of the model that are available in the caret package.

TIP:

There are additional tutorials on the caret package in the optional resources.

1. **[Create Predictive Model Using k-Nearest-Neighbor](https://ut.daacertificate.com/mc/poa?productID=2655&taskID=3351" \l "collapsepoa3326)**
2. **Import and familiarize yourself with the training set.** The file ***Survey\_Responses\_Complete*** is an CSV file that includes answers to all the questions in the market research survey, including which brand of computer products the customer prefers. You will be using this data set to train and test the classifier to make predictions. You will notice that there is a mix of numeric and nominal values; some might need to be converted to factors. Consult the *Survey Coding Key* tab in***Survey Key and Complete Responses*** to gain an understanding of what the survey response values mean. The coding key’s primary function is to help you understand the coding of the survey – you only need the coding key to prepare the labels in your final preferences graph.
3. **Using createDataPartition create training and testing sets.** These will be created from the *Survey Responses Complete.csv* file. The training data should represent 75% of the total data and the remaining 25% will be used for testing. After optimizing your model you'll later use it to make predictions on the incomplete surveys.
4. **Run the KNN classifier on the training set with 10-fold cross validation**. In the task one, you used KNN to predict a specific *numeric*value (the sales volume of a new product); in this task you will use the KNN classifier to predict *nominal* data (a customer's computer brand preference). The main difference is that, in numeric prediction, you are inferring a real number (such as how many products may be sold in a period of time) while, in nominal (or categorical) prediction, you are selecting what class a given observation belongs to. Remember that the data mining algorithms make the prediction in both types of tasks based on the similarities and differences between the attributes (columns) of observations (rows). You will be classifying "brand" in this task. After running KNN:
   1. Assess the performance of the trained model and record the Accuracy and Kappa scores for each K value the model used during training.
5. **Make predictions.**Using the KNN model you just built and the testing set created previously make predictions using the predict() function.
   1. After making the predictions using the testing set use postResample() to assess the metrics of the new predictive model.
      1. Assess the performance of the predictive model and record the Accuracy and Kappa scores.

#### [4. Create Predictive Model Using Decision Tree Classifiers](https://ut.daacertificate.com/mc/poa?productID=2655&taskID=3351" \l "collapsepoa3327)

1. **Run the Random Forest classifier on the training set with 10-fold cross validation**. Using the same caret pipeline you developed for the previous task and the same training and test sets build a trained model using the Random Forest classifier in the caret package. After building the trained model:
   1. Assess the performance of the trained model and record the Accuracy and Kappa scores for each number of features the model used during training.
2. **Make predictions.**Using the trained model you just built and the testing set created previously make predictions using the predict() function.
   1. After making the predictions using the testing set use postResample() to assess the metrics of the new predictive model.
   2. Assess the performance of the predictive model and record the Accuracy and Kappa scores.

#### [Predict Product Preference](https://ut.daacertificate.com/mc/poa?productID=2655&taskID=3351" \l "collapsepoa3328)

1. **Compare optimized predictive models and select one.** Compare the error metrics for your optimized classifier with the error metrics for your optimized Decision Tree or Nearest Neighbor model. Select which predictive model performed better on this data set.
2. **Use the test set (incomplete responses.csv) and apply the predictive model you selected (KNN or Random Forest).**
3. **Use the predict() function and your trained model to make predictions against the testing set**. Be sure to record the performance metrics for your report.
4. **Create a simple chart in Excel** to clearly illustrate the predicted brand preferences.
5. **Write your Customer Brand Preferences Report.** Write a one or two page summary of your methods and findings, including the chart you prepared and collect the results that you saved.
6. **https://s3.amazonaws.com/gbstool/pub/images/mentor_review.pngSubmit your Customer Brand Preferences Report.** Create a Zip file that includes your report and R results. Click on **SUBMIT CUSTOMER BRAND PREFERENCES REPORT** in the **Submit Your Work** tab and upload the document.