**Data Scientist Guidelines for Breakout Sessions**

**Roles:**

**Lead Data Scientist**

Lead discussion

Guide work

**Support Data Scientist**

Run code, models

Cut and paste output to the slides at the end of the session

**Session Outline**:

1. Remember you have 2 hours total.
2. Discuss possible features, copy/paste/run models, review performance
3. Make sure to **leave 15 - 20 minutes at the end** for slide finalization
4. **Extra Time**:
   1. Examine feature importance through H2O
   2. Examine partial dependence plots (see Interactive with H2O below)
   3. Discuss what kinds of features could be added to the model that aren’t here
   4. Discuss what difficulties there might be for incorporating this kind of model into the company’s current workflow. Why? What are some solutions?
   5. Discuss what a “perfect” solution would look like for employee retention.
   6. Include notes of these discussion on the Future Work slide.

**Workflow:**

1. Run all code in file “**10-import-hr-data.Rmd**”
2. Run all code in **”30-feature-engineering.Rmd**”
3. Run all code in file “**40-modeling.Rmd**”
   1. Look at **model output**, specifically how much money the model would have saved and how many employees would have been retained by using the model.
   2. Look at the “**scaled variable importance**” to understand what variables are the most important in the model. They’re scaled so we can compare them to each other, otherwise units would be a problem.
4. Consider **feature engineering** and how you might improve the model. Specifically, what information (that is, features) would help the model make more “informed” decisions?
   1. **ADDING FEATURES**
      1. Go to file ’**30-feature-engineering.Rmd**”,
      2. Find the feature you want
      3. Copy and paste the name of the feature you want into your pipeline in file 40-modeling.Rmd.
      4. Paste **after ‘*attrition %>%***’ but **before** ‘***select(everything())*** *%>%*’. Make sure to include “**()**” after the feature name because the features are functions.
         1. **Example**: *attrition %>% monthly\_income() %>% select(everything())*
         2. This would include the feature monthly income
   2. **REMOVING FEATURES**
      1. In the ‘*select(everything())*’ function, include the name of the feature you wish to remove by typing its name with a ‘-‘ (minus sign) in front.
         1. **Example**: ‘*select(-monthly\_income, everything())’*
         2. This would remove variable created by the feature “monthly\_income()”.
5. Repeat **Step 2**.
6. Repeat **Step 3.**
7. Finalize slides
   1. Final model performance
   2. Important features
   3. Future work
   4. Interesting partial dependence plots  
      (note that you’ll need to use screen clipping to get the graph into Google Slides)

**Features that are not pre-coded in file 30**

1. Discuss the idea with your group (keeping an eye on time), and think about **why** it might help, **how** you could include it in the model, and **potential difficulties** for incorporating it.
2. Go to file **40-modeling.Rmd**
3. Scroll down to the **## Future Work** section
4. Briefly record and summarize the idea / discussion here

**Interacting with H2O**:

You can use H2O to look at feature importance, and to examine partial dependence plots.

Feature importance gives you a high-level view of the relative importance of a feature (e.g. a is twice as important as b)

Partial dependence plots answer the question, “All else being held equal, what is the effect of different levels of this feature?”

1. Open up a browser, and type:[**http://localhost:54321**](http://localhost:54321)
2. At the top of the screen, click on “score”, then “partial dependence plots”
3. The model name is: **rf\_fit2**
4. Frame is the name of data frame: **attrition\_fe**
   1. The frame with the highest value extension is your most recent data frame
      1. Example: attrition\_fe…\_61 is after attrition\_fe…\_50, so attrition\_fe…61 is the most recent data frame loaded into H20.
5. Click “compute”
6. H20 shows the partial dependence plots for the top 10 most influential predictors in the model.
   1. Partial dependence plots are what the model predicts will happen when only 1 variable changes while all others are held constant. It’s a way to look at how much of an effect changing a single variable can have the outcome (in this case, the chance of an employee staying at the company.)

**Keyboard shortcuts for RStudio**:

Mac: **shift + option + k** (to exit the viewer, hit **shift**)

Win: **shift + alt + k** (to exit the viewer, hit **shift**)

In RStudio, click “**Help**” at the top of the screen, then click **“Keyboard Shortcuts Help**”

**Helpful keyboard shortcuts for running and commenting code**

1. Run everything in R Notebook
   1. Mac: **option + command + r**
   2. Win: **ctrl + alt + r**
2. Run current chunk
   1. Mac: **shift + command + return**
   2. **Win: shift + ctrl + enter**
3. Run selected lines:
   1. Mac: **command + return**
   2. Win: **ctrl + enter**
4. Comment out an entire line:
   1. Mac: **command + shift + c**
   2. Win: **ctrl + shift + c**