DISCUSSION POINTS/AGENDA:

Here are tasks that we can divide amongst ourselves, decisions that we can collectively make, and points that we should discuss. We should come to a firm conclusion on each issue if we can do so, identify next steps to come to another conclusion, and divvy out tasks that are separable. We should also identify deliverable dates.

**TASK 1:** MERGING THE DATASET. WHOEVER MERGES THE DATASET MUST PROVIDE IT TO THE REST OF US BY A DATE CERTAIN. WE SHOULD ALL AGREE ON THE DATE.

1. DEALING WITH THE DIFFERENT DATASETS
   1. Merging
      1. Countries
         1. The 2nd-place winner added this to his dataset, and some of these variables may have predictive value, such as longitude/latitude and language spoken. After all, of the users are in the USA. More information differentiating the destination countries may be helpful, rather than redundant.
      2. Age/Gender Buckets
         1. Although the 2nd-place winner added this to his dataset, it is unclear why.
      3. Sessions
         1. The sessions dataset can be joined on the id column. The id column consists of the user ids in the training dataset.
         2. One solution to deal with the sessions dataset is to collapse the information – id, summary of elapsed time, and number of actions.
         3. Another solution is to conduct PCA, Ridge regression, Lasso Regression, or analyze feature importance (for example, using a boosted trees model) on the sessions dataset to determine which features are most important to predicting elapsed time.
         4. Extremely basic EDA of the sessions dataset will probably be useful for our presentation. What types of actions were most common? What devices were used most commonly?

**TASK 2:** PRODUCING A CORRELATION PLOT, A MISSINGNESS CHART, AND CONDUCTING PCA.

1. MISSINGNESS
   1. VIM chart
      1. What type of missingness do we have?
      2. What is our solution? (Imputation?)
2. DIMENSION REDUCTION/FEATURE SELECTION
   1. Correlation plot
      1. Do we have any multicollinearity? Any theories as to why?
   2. PCA
      1. How many components do we have?
      2. How much variance in the dataset does each component explain/account for?
      3. What exactly have we sacrificed in interpretability, given our prior analysis?

AFTER THESE TWO STEPS ARE DONE, THE PRINCIPAL COMPONENTS SHOULD BE PLACED INTO A REDUCED MODEL. ONCE WE HAVE A REDUCED MODEL WE CAN DECIDE ON:

**TASK 3:** ADDING FEATURES TO (HOPEFULLY) IMPROVE PERFORMANCE

1. FEATURE ENGINEERING
   1. Time-lag variables
      1. Do we want to include them? What is the rationale for their inclusion, other than that the 2nd-place winner included them and won?
   2. Stacking (add a column of predictions for a country destination/NDF)
      1. We should discuss which model type might be best for this purpose.
         1. Support Vector Machines.
         2. Random Forests.
         3. Other models.

**TASK 4:** THE ENTIRE GROUP RUNS MODELS ON THE REDUCED DATASET.

1. MODEL SELECTION (Each group member can run a model)
   1. Logistic Models
      1. Logistic Regression
      2. Neural Network/Linear Discriminant Analysis
      3. Note: we do not need dummy variables to run these models.
         1. They can take PCA components as features.
         2. They can take continuous variables as features.
   2. Tree-Based Models
      1. GBM
      2. XGBoost/AdaBoost
      3. Note: Tree-based models are modeling decision splits.
         1. They can handle categorical variables.
         2. They can handle categorical variables simultaneously with continuous variables.
   3. Support Vector Machines
      1. Perhaps this model type can provide the predictions for stacking.

**TASK 5:** A DISCUSSION OF FINDINGS AND THE PRODUCTION OF A REPORT/PRESENTATION.

1. INITIAL IDEAS FOR VISUALIZING THE FINDINGS OF THE REPORT/PRESENTATION:
   1. A predictive Shiny app.
      1. What precisely is the app predicting?
      2. What does the app add to our findings?
   2. A map.
      1. What is the map of?
      2. How does the map illustrate our findings?