**GOING FORWARD**

This document is intended to memorialize what we have agreed to as a group for next steps.

Task 1 – Merging the Relevant Datasets – Yannick – **Due by Friday, June 17th, 2016**

1. Countries dataset. This will be joined on the countries variable to the train and test.
2. Age/Gender Buckets. This will be joined on the age variable; the age values in the training set and test set will be transformed into age buckets.
   1. If this produces additional missing values, that should be noted.
3. Sessions. This will be joined on the user id/id. Additionally, we will collapse this dataset into:
   1. The number of actions;
   2. The total seconds of elapsed time;
4. Simple graphical/numerical EDA on:
   1. The types of actions and the types of devices (in sessions)
   2. The date account created and the first activity. (How long does it take people to book?)

Yannick will then hand the dataset over to Michael.

Task 2 – Missing Data Analysis, Multicollinearity, and PCA – Michael – **Due by Monday, June 20th**

1. Produce a Missing Data Chart in VIM(aggr)
   1. Potential solutions to missingness (Imputation/KNN)
2. Correlation Plot
3. PCA
4. Creating the reduced dataset.
   1. This requires
      1. Creating a dataset of the components; and
      2. *Transforming the test data set* so that our models can predict from the reduced train to the reduced test.
      3. This dataset will have however many components, plus a date of first booking variable and a response variable (country of destination).

Michael will then hand the reduced dataset over to Zi, who will use Yannick’s dataset to add features to it.

Task 3 – Feature Engineering – Zi – **Due by Wednesday, June 22nd**

1. Time-lag variables:
   1. The difference between the date account created and the first booking. This consists of 4 categories:
      1. 0 (booked immediately), negative (booked before creating account), positive (waited to book), no booking (NB).
   2. The difference between the timestamp of first activity and the date of first booking. This consists of:
      1. 0 (booked on the first active day), non-zero (booked after first active day), NB (no booking).
      2. Note that the date of first booking variable is not in the test data set.
      3. Therefore, we have to use the components in the reduced training dataset to predict the date of first booking on the reduced training dataset, and then add those predictions as a feature to the reduced/transformed test dataset.
   3. The predictions of the date of first booking added to the test dataset will then be used to create lag variables a and b in the test dataset.
2. Additional features may be created.

Task 4 – THE ENTIRE GROUP RUNS MODELS – **Due by Thursday, June 23rd**

Logistic Regression

Linear Discriminant Analysis

GBM

XGBoost

Task 5 – Ensembling – Rob**, Due by Friday, June 24th**

* Rob should probably start researching how to do this now.

Task 6 – Preparing the Report/Presentation – **Due by Sunday, June 26th**

Task 7 – Developing the Predictive Shiny App and Writing the Blog Post – **Due by July 1st, 2016**

The app will consist of a predictive map:

1. 12 countries, highlighted
2. Click on a country to see:
   1. Summary stats from countries.csv
   2. Age/gender stats from age/gender.bkts
3. Input:
   1. Gender
   2. Age
   3. Language
   4. Other (??)
4. Output:
   1. Country you are most likely to visit