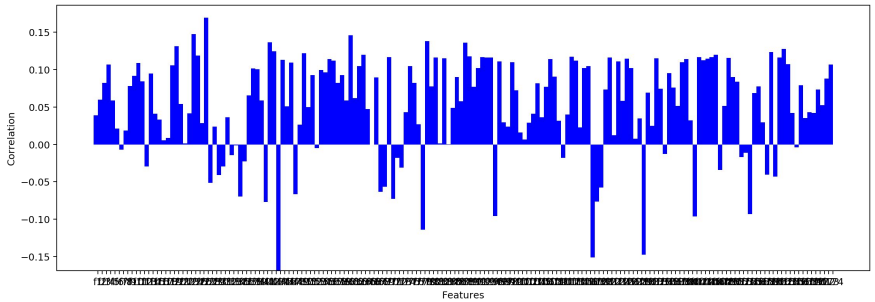
DataFest 2017

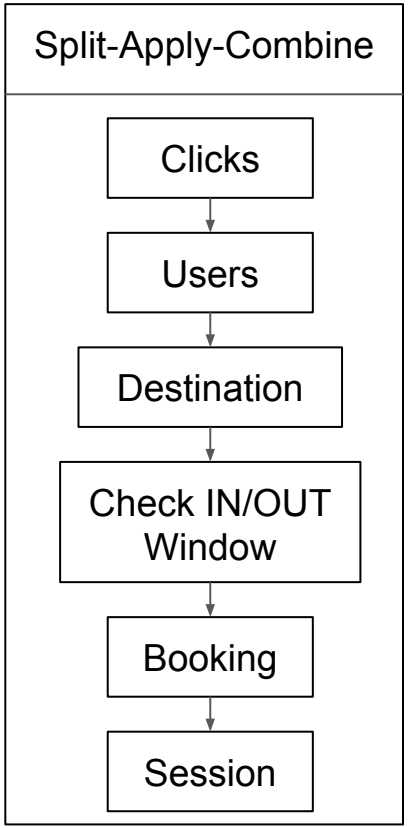
Vizards

Alex Fedorov, Eswar Damaraju, Rogers F Silva

Hypothesis testing with Machine Learning

- **Data**
  - 42317(33971) from 10000 random users
  - Splitting 67% for Train and Validtion and 33% Test
- **Features**
  - Hist price band [VL, L, M, H, VH]
  - Popularity band [VL, L, M, H, VH])
  - Distance band [VC, C, M, F, VF])
  - Hotel Start Rating [0, 1, 2, 3, 4, 5]
  - Hotel Brand [0, 1]
  - Mobile vestion [0, 1]
  - Package [0, 1]
  - Number of hotels is looked
  - Adults
  - Childrens
  - Number of Rooms
  - Location Latitude
  - Location Longitude
  - Destination Distance
  - All Destination Scores

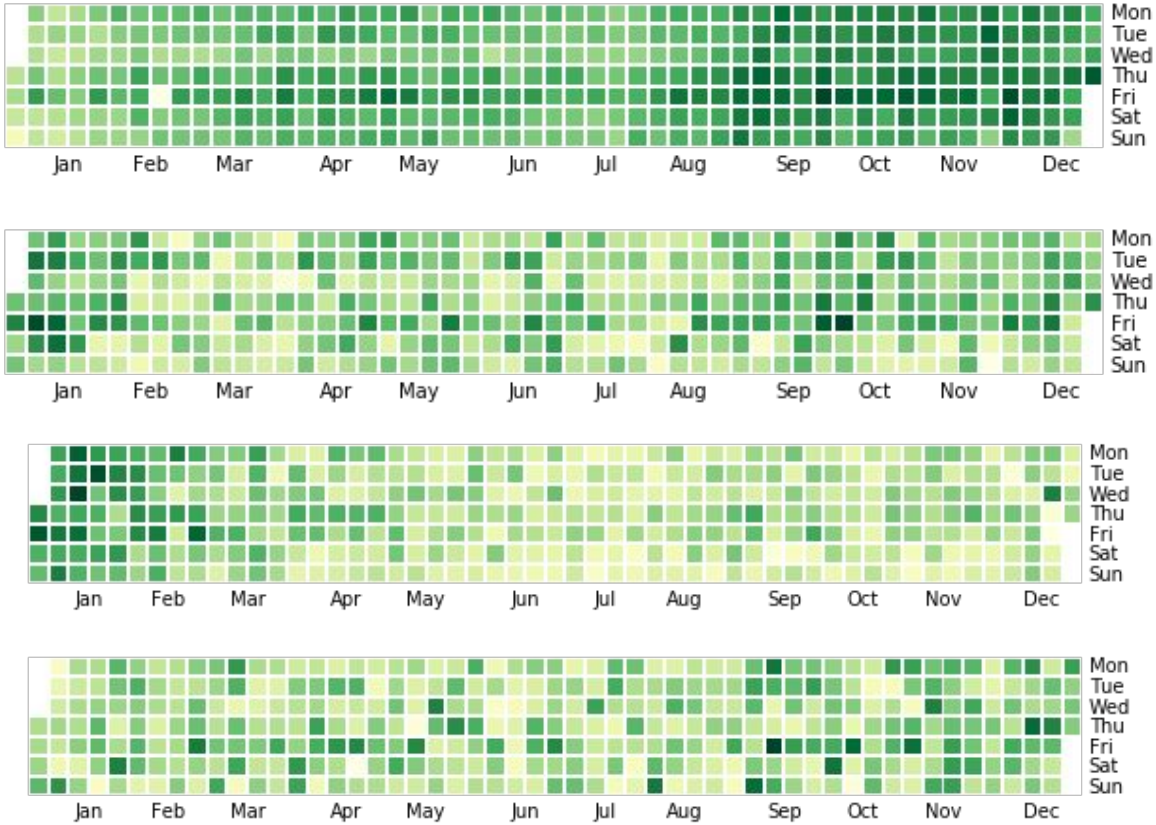




- **Metric - F1 Score**
- **Models**
  - Logistic Regression(0.51, Random Forest(0.59), XGBoost(0.69)
- **Some insights based on Feature Importance of XGBboost (TOP-10)**
  - Destination Distance
  - Location Latitude
  - Location Longitude
  - Number of hotels is looked
  - Hotel Brand
  - Desktop
  - Very High Popularity Brand
  - Number of Adults
  - Star rating 4
  - Package

Booking Ratio - Countries with most users

$$BCR = \frac{Booked}{Booked + Click}$$



USA

Canada

Germany

Mexico