Expedia Hotel Recommendation

Anandan Balaji
17 June 2016

Overview

- Problem Definition
- Dataset
- EDA / Model
- Manual Feature Engineering
- Results

Problem Definition

Predict the hotel cluster for an Expedia customer.

- Hotels are grouped on many parameters and is called **hotel cluster**.
- The hotel clusters are numbered from 1 to 100.

Dataset

The following are the datasets provided and they are available at Kaggle, www.kaggle.com/c/expedia-hotel-recommendations/data

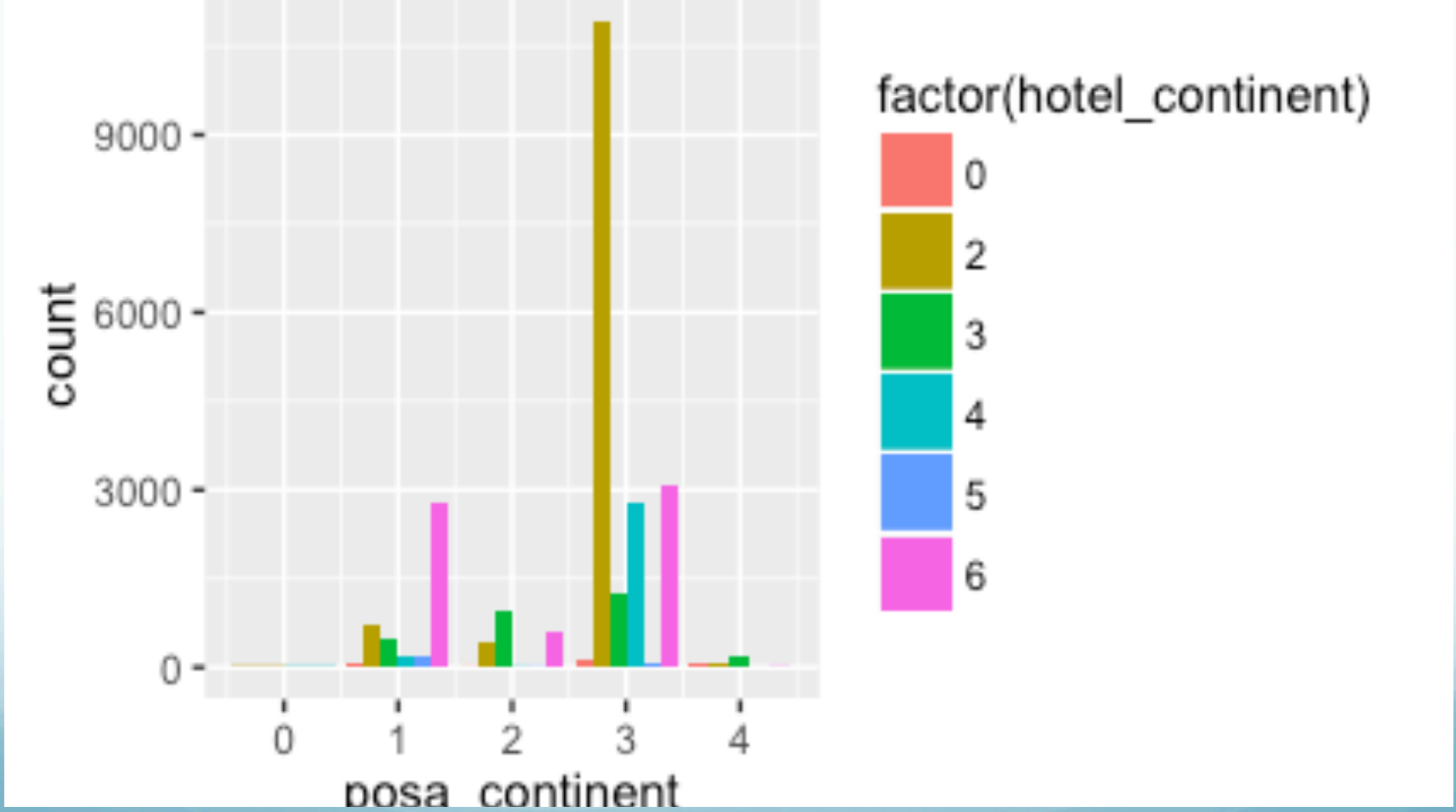
- **train.csv** - the training dataset
- **test.csv** - the test dataset
- **destinations.csv** - hotel search latent attributes
- **sample_submission.csv** - the sample submission file in the correct format

Dataset – Important Variables

Variable	Description
posa_continent	ID of the continent
user_location_contry	Country ID where customer is located
user_location_region	Region ID where customer is located
user_location_city	City ID where customer is located
orig_destination_distance	Physical distance between the hotel and customer at the time of search
srch_destination_id	ID of the destination hotel
hotel_continent	Hotel continent
hotel_country	Hotel Country
hotel_country	Hotel Country
is_booking	1 if a booking, o if a click
hotel_cluster	ID of the hotel cluster

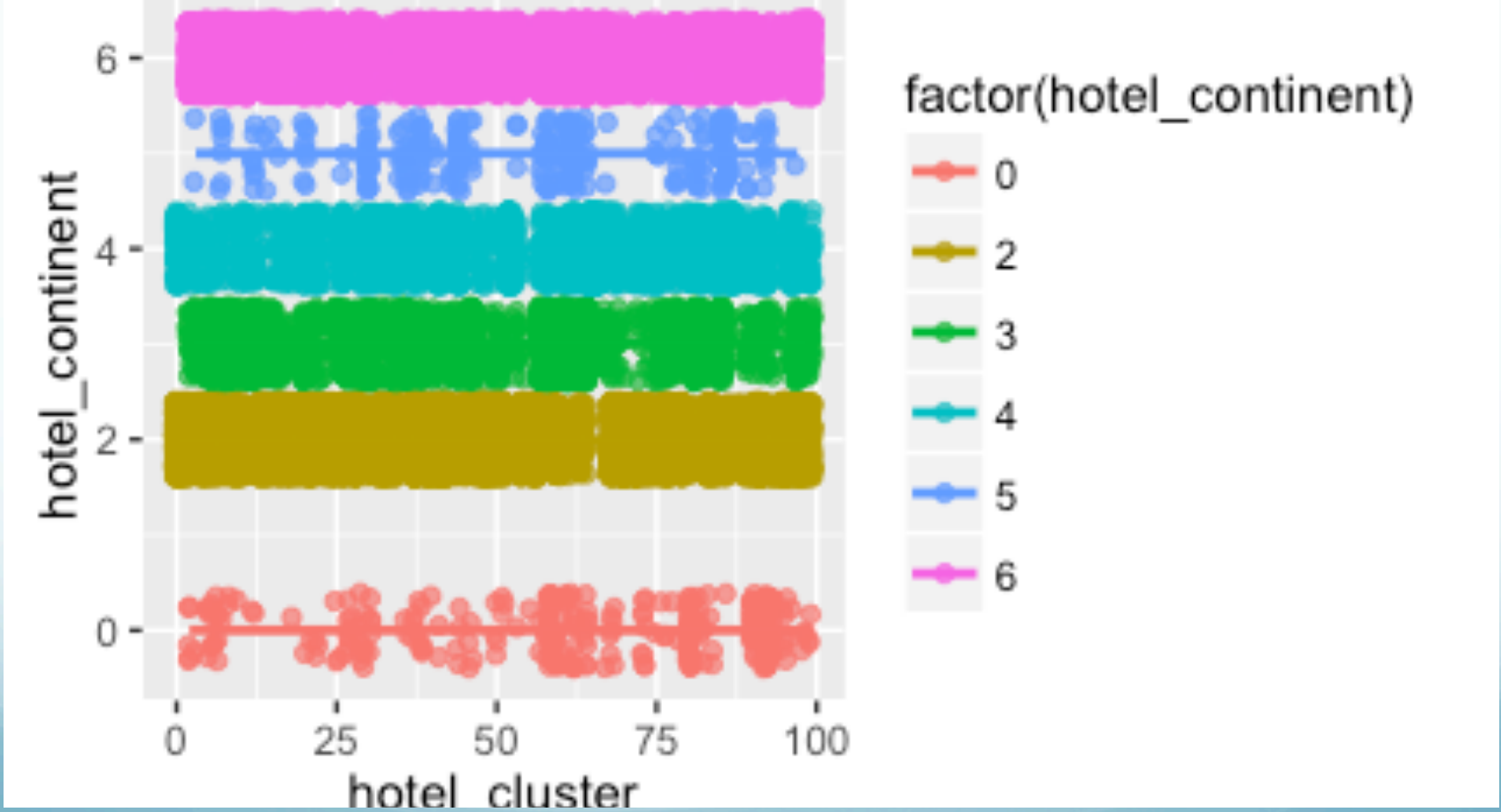
EDA

Source and Destination continent spread



EDA (2)

Hotel Cluster spread



EDA, Model

- By plotting, no pattern/ correlation was found between the dependent and independent variables.

- Linear Regression Model
  - The R-squared value was negligible = 0.005
  - However, the dependent variables **is_booking**, and **is_package** had coefficients which were significant .

EDA, Model (2)

- CART (Classification And Regression Tree)
  - Unable to build tree beyond root node
  - The CP (Complexity factor) was negligible 0.006

- The detailed analysis report is available at https://github.com/abalaji-blr/CapstoneProject/tree/master/Deliverables/ExpediaHotelReco.pdf.

Manual Feature Engineering

- The basic Machine Language algorithms were not suitable for this problem.
- However, we have identified some dependent variables which are significant.
- Let's derive features using them and predict the hotel cluster.

Feature #1

Identify often used hotel clusters

- For a given destination, identify the often used top five hotel clusters
- Also, give importance to **is_booking**
  - If is_booking is 1, give weightage as 1
  - If is_booking is 0, give weightage as 0.15

Feature #2

Use **orig_destination_distance**

- There are few records match between test and training dataset based on orig_destination_distance
- Predict top five clusters using that
- Give preference to this feature result as they are appropriate match when compared with feature 1 results.

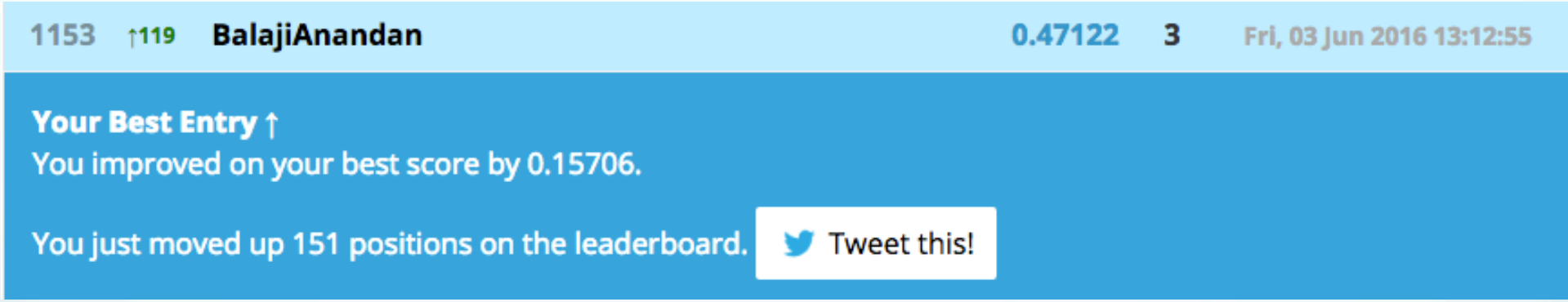
Results

- Combine results from Feature #1 and Feature #2
  - Pick 5 hotel clusters
  - Make sure they are unique.

- The complete R script is available at this https://github.com/abalaji-blr/CapstoneProject/tree/master/Deliverables/ExpediaScript.R

Results (2)

- The Manual Feature Engineering approach yielded the Mean Average Precision Score at 5 (MAP@5) of 0.47122!



Future Work

- Explore advanced Machine Language algorithms like Random Forest, XGBoost etc.

Thank You