Airbnb first booking country classification – Project Summary

Songlin Qing

Problem definition:

- Based on user demographics and browsing session information, predict the first booking country the user will make.

Data description:

- https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data

STAGE 1 – Data subset

- The original dataset is too big and model building is too slow, thus I took 10% of the original dataset, and divided it into train and test set (75%, 25%)

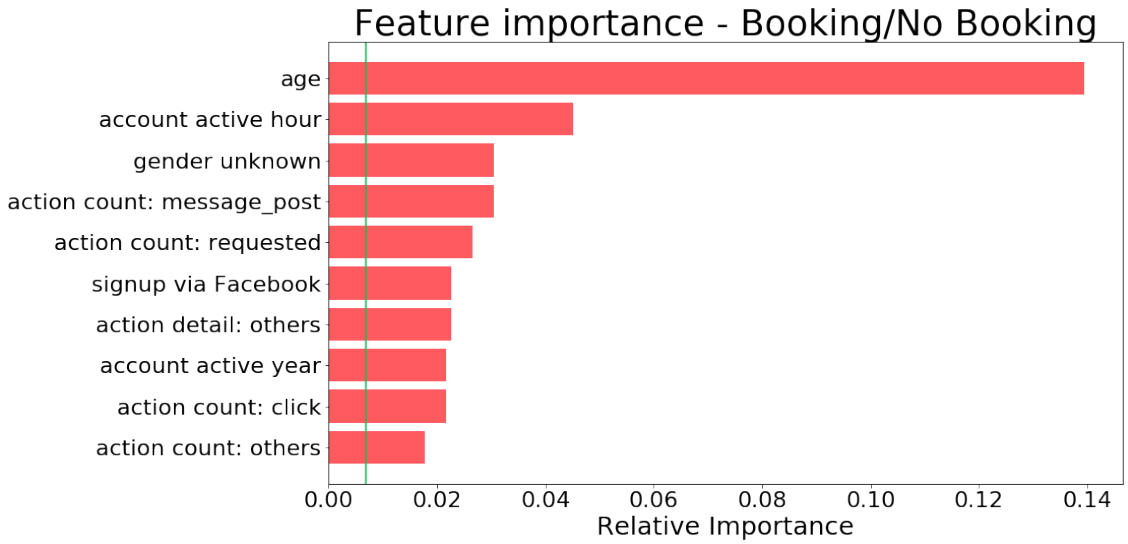
STAGE 2 – Data cleaning and preprocessing

- 42.5% of the age is missing or wrong (<18 or >100)
  Action taken: Different imputation techniques were used: 1. Treat age as a dependent variable and use the other variables to build a regression model; 2. Use MICE package; 3. Use categorical mode age based on some other categories. All three techniques had similar RMSE (~11) thus the simplest technique (3rd) was used.
- 10% of the browser action type and action details are missing
  Action taken: Used categorical mode from the same action category.
- 4% of the session duration is more than 24 hours
  Action taken: Those session durations were clipped to 24 hours.
- 400+ categorical values
  Action taken: rare values that appear less than 0.5% of the total population are replaced with catch-all, to bring down the total number of categorical values to 100+.
- Timestamps were broken down into year, month, and hour
- Browser session durations for each user were broken down to mean, median, standard deviation, skew, kurtosis, and inter quantile range.

Stage 3 – Build models

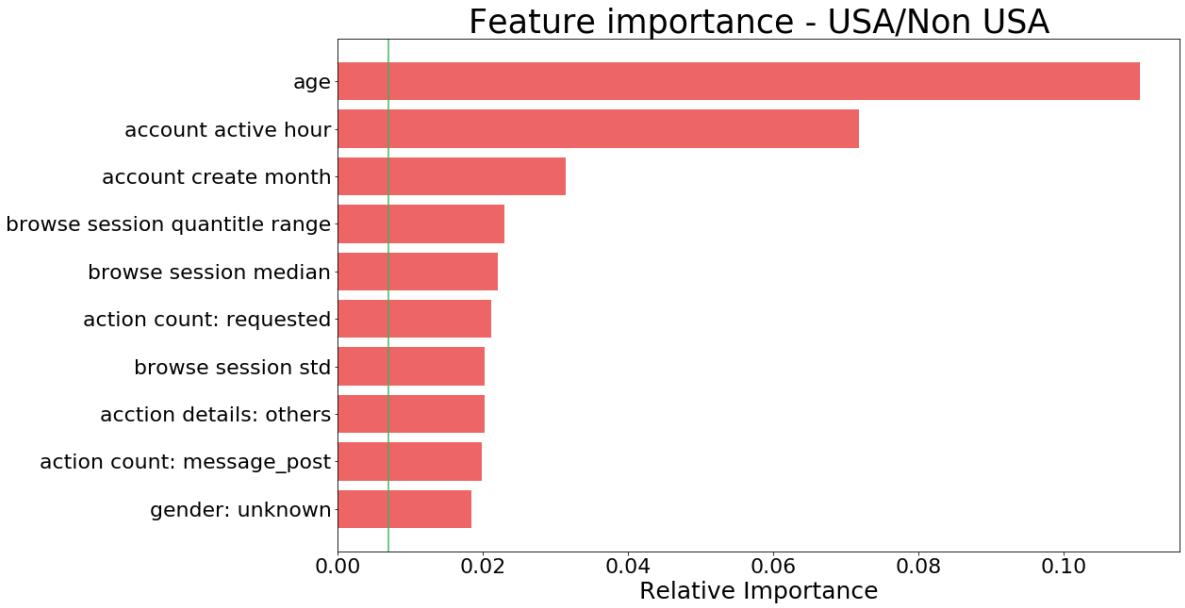
Model 1 – Booking/No-booking

- Destination country was categorized as booking and no-booking.
- Variables were normalized, before feeding into a GridSearchCV pipeline with 5 models: KNN, logistic regression, tree, random forest and xgboost.
- Best model turned out to be xgboost with an training accuracy of 0.71.
- A soft prediction was made for likelihood of booking.
- Most important features are listed below. The green line represents the average feature importance if all variables share the same importance.



Model 2 – US/Non-US

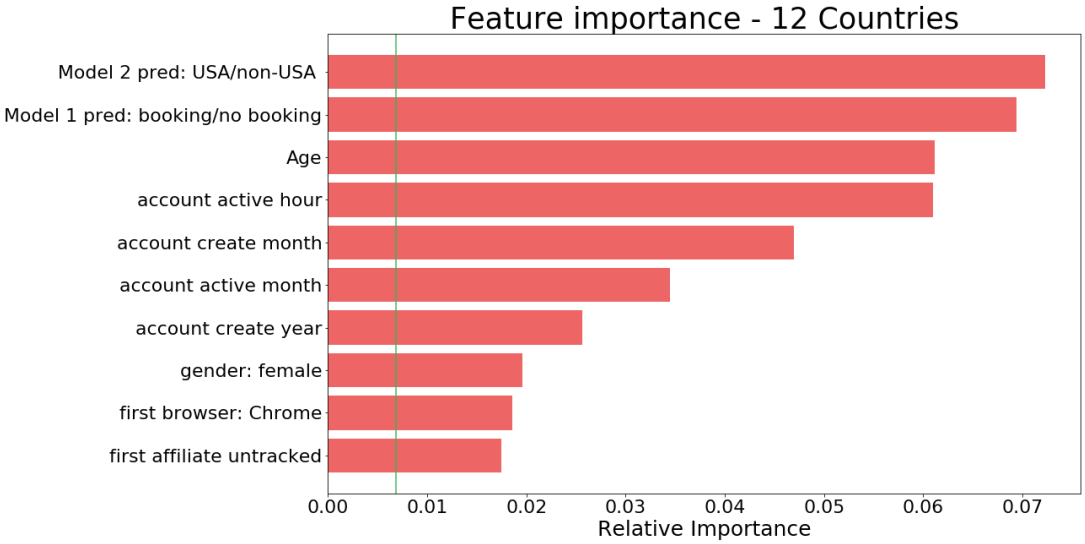
- Destination country was categorized as US and non-US.
- Variables were normalized, before feeding into a GridSearchCV pipeline with 5 models: KNN, logistic regression, tree, random forest and xgboost.
- Best model turned out to be xgboost with an training accuracy of 0.74.
- A soft prediction was made for likelihood of US as destination country.
- Most important features are listed below. The green line represents the average feature importance if all variables share the same importance.



Model 3 – 12 destination countries

- Soft predictions from model 1 & 2 were added as additional features.

- Input data were oversampled. Original plan was to oversample every country to the same level. However, due to the huge difference between no booking, US and the rest of countries, the result dataset became huge and took long time to train. Instead, I oversampled the countries (except US) to the same level of others (5%).
- Variables were normalized, before feeding into a GridSearchCV pipeline with 2 models: logistic regression and xgboost.
- Best model turned out to be xgboost with a training accuracy of 0.71.
- Most important features are listed below. The green line represents the average feature importance if all variables share the same importance.



Stage 4 – Validate result

The models were applied to test data to validate the model performance. The first two models performed relatively well, with accuracy of 0.697 and 0.718. However, the last model performed poorly with accuracy of 0.63. The learning curve suggest overfitting problem and more data could help to improve the model accuracy.