Airbnb's New User Bookings Kaggle Competition

Lisa Oshita

April 6, 2018

1 About the Competition

Founded in 2008, Airbnb is an online accommodation marketplace featuring millions of houses, rooms and apartments for rent or lease in over 200 countries. As part of a recruiting competition, Airbnb partnered with Kaggle to host a data science competition starting in November 2015 and ending in February 2016. The task of this competition was to build a model to accurately predict where new Airbnb users will make their first bookings. There are a total of 12 possible destinations to predict: US, France, Canada, United Kingdom, Spain, Italy, Portugal, Netherlands, Germany, Australia, Other and no destination found. No destination found indicates that a user did not make a booking, while Other indicates that a booking was made to a country not already listed.

2 XGBoost, Random Forest, Stacked Models

Extreme Gradient Boosting (XGBoost) is a fairly new method of supervised learning that performs consistently better than single-algorithm models. It is a form of gradient boosting that introduces a different, more formal form of regularization to prevent overfitting—enabling it to outperform other models. Additionally, XGBoost algorithms are fast. At its core, the algorithm is parallelizable which allows this model to fully utilize the power of computers. As XGBoosts have been used to place highly in Kaggle competitions and were also used by many of the top 100 participants of this Airbnb competition, this was the model I first decided to explore and implement.

The second model I decided to implement was a random forest. Random forests are another form of ensemble modeling that have performed well in Kaggle competitions.

As the top three winners of this Airbnb competition used a form of stacked modeling, I also decided to explore a general form of this. Stacking, also called meta ensembling, is a technique used to combine information from individual predictive models to create a new model. Stacked models usually outperform its base models—as it's able to correct and build upon the performance of those base models.

3 Exploratory Analysis

imbalanced classes ages gender number of missing values

For this competition, Airbnb provided six data sets for participants to use in model building. I used two of these data sets: train_users, which contains information about the users dating back to 2010, including where they first booked, and sessions, which contains information about user's web session activity. The train_users data contained 213,415 rows and 16 original columns. The sessions data contained 1,048,575 rows and six original columns. To reduce computation time when working with the sessions data, a function was written to sample 10% of the rows belonging to each unique user within the data. The sampled data came out to include 104,424 observations. All following anayliss and model building was done on this sampled sessions data.

Table blah shows the first six rows of the sessions data. The data contained five features describing each user's web session activity: action (193 levels), action_type (9 levels), action_detail (93 levels), device_type (13 leels), and secs_elapsed. There were 12,994 unique users within this data. A majority of the users within this sessions data were also in the train_users data.

Exploring the train_users data revealed several factors about the data.

Percentage of the data	**Destination**
NDF	58.35%
US	29.22%
other	4.73%
FR	2.35%
IT	1.33%
GB	1.09%
ES	1.05%
CA	0.67%
DE	0.50%
NL	0.36%
AU	0.25%
PT	0.10%

4 Feature Engineering

From the train_users data I created a total of ? features. I pulled apart the month, year and day of the week of the date features (date account created, date first active, date of first booking) and created a season variable (winter, spring, summer, fall). I also calculated the difference in days between each date feature: days between the date an account was created and the date of first booking, and days between date first active and date of first booking. I also cleaned the age and gender features.

From the sessions data I created a total of blah features. I aggregated the data by each user and created features counting the number of unique levels for each of these features. From secs_elapsed, I also calculated summary statistics like the mean and median for each unique user. These features were then joined by user id to the training data.

The following is a list of all features included in the training data:

Features in the original data

- signup_method
- signup_flow
- language
- affiliate_channel
- affiliate_provider
- first_affiliate_tracked
- signup_app
- first_device_type
- first_browser

Features derived or cleaned

- Year, day of the week, month and season of date account created, date first active, date first booking
- days between date account created and date of first booking
- days between date first active and date of first booking
- age
- gender
- count features created from the sessions data
- mean, median, standard deviation, minimum, maximum of secs_elapsed

One-hot encoding was used to convert categorical features to a form that works better with machine learning algorithms. Essentially, a boolean column indicating a 1 or 0 was generated for each level of the categorical feature. Continuous features were left as is. After one-hot encoding and feature engineering, there were a total of 596 features to use in the models.

5 Model Building and Results

The processes used for building the three models (XGBoost, random forest, and stacked models) was not linear. It was an iterative process in that, when new findings were discovered or new features were added in, I went back a few steps to make changes. But overall, the same process was used for all three models.

The full data was partitioned into one training set, containing 70% of the full data (149,422 rows), and one test set containing 64,029 rows. All model building was performed on just the training set. For both the XGBoost and random forest models, five fold cross-validation was performed on the training set, including all 596 features. Both models achieved 87% classification accuracy in this process, as well as a Normalized Discounted cumulative gain score of 0.92. However, both models only made predictions for a few out of the 12 possible countries. Predictions made included just the US, Other, and no destination found. Examining feature importance for each model showed that not all features were contributing to the predictions. So, the top 200 most important features for each model were extracted. The models were then refit again to the same data, but only with those top 200 features, and five fold cross-validation was again performed. Accuracy and NDCG scores remained the same, but computational time was much faster. These top 200 features were the only features considered from this point forward.

Country	**N**	**Proportion**
AU	5670	0.05
CA	5000	0.04
DE	5944	0.05
ES	6300	0.05
FR	7034	0.06
GB	6508	0.06
IT	5955	0.05
NDF	35000	0.30
NL	5340	0.05
other	7066	0.06
PT	5320	0.05
US	20000	0.17

To account for the highly imbalanced classes, various techniques were explored—regular oversampling with replacement combined with undersampling from the overrepresented classes, before settling on synthetic minority oversampling techniques (SMOTE). With SMOTE, underrepresented classes are upsampled by generating synthetic examples by selecting neighbors from the k-nearest neighbors of the minority classes. This was combined with an undersampling of the majority class. The resulting training set contained 115,137 observations. Table 5 shows the number of observations and proportion for each country after SMOTE and undersampling was performed.

The same model fitting process (cross-validation with only the top 200 fea-

tures) was again performed on this new training data. For both models, accuracy and NDCG scores remained the same. However, with this new training data, the models were now able to make predictions for all countries instead of just a few. Although the predictions for these minority countries were not significantly accurate. Tables blah and blah show the confusion matrices for the XGBoost and random forest fit to this new training data. Figures blah and blah show the feature importance for each model.

After individually exploring the random forest and XGBoost models, they were combined into one stacked model according to this guideline. The training data was partitiioned into five folds of around the same size. Each model was built on the training folds, and tested on the held out test fold. For example, the XGBoost was built on folds one, two, three and four, and used to make predictions on fold five. In the second iteration, the XGBoost was build on folds two, three, four and five, and used to make predictions on fold one. This process was repeated for both models, until each fold had been used as a test fold. Predictions from these models were stored as columns in the training data. Then, each model was fit to the full training data (ignoring the folds and predictions created in the previous step), and used to predict on the held out test set. Predictions from these models were stored in two columns in that test set. A final XGBoost was used as the model to combine, or stack the information from the previous processes. The XGBoost was fit to the predictions stored in the training set, and tested on the predictions stored in the test set. Resulting accuracy and NDCG score for this stacked model was the same as all previous models.

6 Discussion and Conclusion

Various strategies were explored in an attempt to improve the performance of these models. However, every method explored made no difference in the model's accuracy. One of the biggest reasons these model's hovered at 87% is the highly imbalanced classes. Three destinations—no destination found, other, and US—make up over half of the data. The remaining eight, underrepresented countries each make up less than 5% of the data. Essentially, there is not enough information and data for the models to be able to differentiate and make predictions for those underrepresented countries. However, because those three countries make up such a substantial part of the data, the accuracy for each model remained fairly high. In order to create a model that can make accurate predictions for all of the twelve possible countries, a method for handling these highly imbalanced classes would need to be implemented.

Possibilities for future work on this, one might consider to combine more models of different varieties. The second place winner of this Airbnb competition built numerous models that did not just predict destination, but other models that also predicted missing values of certain features. Those features were then used in predictions for more models. Overall, it appears that the top three participants were combined multiple models in several layers. One might

consider doing this, and expanding or adding on to the stacked model I implemented. One reason this stacked model did not result in a better performance than the two base models, is that stacked models generally work well if the two base models perform in different ways (e.g. if one model predicts one category accurately but not another, and the other model predicts the category that the other model is weak with, accuratley, but not the other way around). In my case, both models could not make accurate predicitons for the underrepresented countries, thus, the stacked model also could not.

7 References