Rounak Banik
IIT Roorkee
rounakbanik@gmail.com
+91 84398 60325

Airbnb New User Bookings

14th August 2017

OVERVIEW

The problem this capstone project aims at solving is predicting where a newly registered Airbnb user will book his/her first travel experience. By analyzing a variety of data such as user informations, country information and user browsing session records, the project aims at coming up with a model that could accurately detect where a user would book his/her first experience from an array of 34000+ cities across 190+ countries.

THE CLIENT

The Client in question is Airbnb, an online marketplace and hospitality service enabling people to lease or rent short-term lodging including vacation rentals, apartment rentals, homestays, hostels beds, or hotel rooms,

Since its inception in 2008, Airbnb has expanded into more than 190 countries and 34,000 cities across the planet. It has disrupted the hotel industry and has become the premier choice of booking for many travelers across the globe with people booking a wide variety of spaces from tree houses to houseboats to apartments.

When a user first signs up for Airbnb, it is in the user's and Airbnb's best interests to show homes and places in a city and country that the user intends on visiting. Since it is not possible to explicitly ask the user for this data and expect the user to update his/her preference (a user may just be browsing, s/he may change preference from time to time and may not update that in his/her profile, etc.), it would make sense to actually build a predictive model around the user's browsing data and give predictions based on it.

This would serve the following purposes:

1. Share more personalized content with community
2. Decrease the average time of booking
3. Better forecast demand

In other words, based on the analysis done and the model built, Airbnb will be able to provide a much better user experience by offering a news feed tailored to each user's preferences for their next destination.

This, in turn, would decrease the time the user took to book their first host as they will spend considerably less time searching for homes and more time comparing appropriate options.

Finally, it would also help Airbnb which cities and countries are in hot demand at a particular time of the year. This would help them in suggesting more competitive pricing to hosts in these regions. Additionally, it would also help the client focus their marketing efforts in a region where demand is high, leading to better sales.

THE DATA

The data is already available to us in the form of a Kaggle competition hosted by Airbnb in 2015. Therefore, there isn't a need for additional data mining or web scraping. The data will however have to be cleaned and wrangled before any analysis is performed on it.

The data provided by Airbnb is in the form of CSV files and are listed below:

1. **train_users.csv**: The training set of users
2. **test_users.csv**: The test set of users. Contains user information such as gender, age, language, signup and device information
3. **sessions.csv**: Web sessions log for users. Contains time, type and details of various user actions.
4. **countries.csv**: Summary statistics of destination countries in this dataset and their locations
5. **age_gender_bkts.csv**: Summary statistics of users' age group, gender, country of destination
6. **sample_submission.csv**: Correct format for submitting predictions

APPROACH AND MILESTONES

The approach to solving this problem is subject to change as I progress with the Career Track and learn new concepts and approaches. The tentative broad overview of solving the problem is explained below.

Data Wrangling

The first step would be to load the given datasets into Pandas dataframes and clean them. This would be followed by various wrangling methods to arrive at data on which analysis and prediction can be performed.

Data Visualisation

The second step is to create visualisations for the cleaned data and try to come up with a few preliminary hypothesis. The graphs, charts and other visualisations will also be a very important part in creating our story for the project.

Statistical Analysis

Based on the hypothesis formed in the previous step, the next step would be to perform various statistical analysis on it to test if the hypothesis is indeed correct. A wide array of tools such as regression analysis, scatter plot conclusions, chi square significance tests, z statistics and t statistics will be employed in this step.

Machine Learning

The next step would be to build a predictive model using elementary ML methods and, if necessary, deep learning.
The model will be incrementally improved by testing it against the test users dataframe created from one of the Airbnb datasets.

Data Story, Results and Conclusion

The final step is to report the results of the analysis performed and the accuracy of the predictive model. This step will involve creating a story around the initial problem, the problems it aims at solving and the insights gained from the data. This will be followed by explaining the intuitions involved in building the ML model, the incremental improvements involved, the accuracy and future prospects of improvement.

DELIVERABLES

The following should be considered as deliverables as part of the project:

1. **Jupyter Notebook:** Contains all the code involved as part of wrangling, analysis and building predictive models.
2. **Project Report:** A Document highlighting the entire process of the project.
3. **Presentation:** A Slide Deck to be presented to the clients as the final product of the analysis performed and the model built.