Airbnb New User Bookings
Capstone Project Proposal

Mohammed Shinoy
mshinoy@uwaterloo.ca

Abstract

The project proposal is to create an ML model to Predict the first booking of a new customer on Airbnb based on previous user data containing demographics and sessions data. This information may be crucial to the company to provide relevant recommendation and support for the customer.

Keywords: Airbnb, Supervised Learning, sci-kit learn

1. Domain Background

The customer is the focus of any business, keeping them happy and engaged will benefit both the customer and the business equally. Therefore it pays to know your customer's needs as it helps the business keep inventory, support, and recommendations on hand just before the customer asks for it. This concept is implemented by many technology companies and industries to provide recommendations to the consumer before they even think about it. Airbnb being a revolutionary technology company which has disrupted the short rental space is no exception, It would be helpful to know the user's behavior in order to efficiently cater their products and services. The motivation for this project stems from the fact that vast amounts of data are collected from users on a daily basis by many companies. They rely on prediction to serve their customers better. A classification algorithm can be used for such purposes. Data cleaning, Data wrangling, and visualizations can be performed for better analysis and improve the systems that are currently in place. All in all an effective learning experience.

2. Problem Statement

Airbnb is a trusted community marketplace for people to list, discover, and book unique accommodations around the world online or from a mobile phone or tablet [1]. Utilizing prior user data effectively can help Airbnb drive up user experience, customer retention and customer satisfaction.

The needs of a new user can be predicted based on the demographic data of all the previous users. The problem statement is to predict the location a new user would book based on demographics and session data. User data contains the date, device type, geographical data etc. Based on this given data, create a machine learning model that will be enable accurate prediction of a new user booking, given the users location, device type etc.

3. Datasets and Inputs

The dataset contains 5 .csv files with information necessary to make a prediction. They are:

- countries.csv - summary statistics of destination countries in this dataset and their locations
  - Number of rows : 10
  - Number of columns : 7
  - Might not be a highly relevant dataset as the data it contains can be dropped, especially distance wont affect the prediction as the users in question are all from US and all countries are at the same distance for every user. Further analyses required
- age_gender_bkts.csv - summary statistics of users age group, gender, country of destination
  - Number of rows : 420
  - Number of columns : 5
  - Might not be relevant dataset as most of this data is obtained in the train_user.csv such as age, gender, country etc.Further analyses required
- train_users.csv - the training set of users
  - Length of rows : 213451 rows
  - Number of columns : 16
  - Highly relevant data as we are supposed predict based on these columns, contains user id, dates of account creation, first booking dates, gender, age, signup method, signup app, destination etc.
- test_users.csv - the test set of users
  - Number of rows : 62096
  - Number of columns : 15
  - test_users.csv has the same data columns as train_users.csv but not the destination and date first booking as they are users. We have to predict the destination for this set.
- sessions.csv - web sessions log for users
  - Length of rows : 10567737
  - Number of columns : 6
  - Information regarding every user session with respect to the actions taken, device used and secs_elapsed being some of the important columns.

As this was a Kaggle competition. The dataset is provided by Kaggle [2] and Airbnb. They can be obtained here.

4. Solution Statement

Most probably there will be similarities in user behavior for people in the same demographics. These similarities will be helpful in creating an ML model using supervised learning to predict new user behavior. ML techniques such as SVM, Random Forest, AdaBoost etc along with GridSearchCV can be used to model an algorithm which would provide and optimal result.

5. Benchmark Model

As this is a Kaggle competition a benchmark model would be the best Kaggle score for the test set, which comes in at .88697 NDCG score. If the model trained comes above 0.86 NDCG the

model can be deemed useful and ready.

The test set for this model is provided by Kaggle as a dataset. The testing will be done on this set and the benchmark model is hosted by Kaggle with which we can compare our model performance. A personal goal would be to be in the top 20% ie above 0.88183 NDCG score of the Kaggle Private Leaderboard.

6. Evaluation Metric

The model prediction for this problem can be evaluated in several ways. Accuracy being the basic one but this has known shortcomings. A good evaluation metric would be f1_score. This would allow us to get a sense of model performance if the dataset tested is skewed. This is also a good metric for keeping a check on false positives and negatives.

The official evaluation of this project is done by Kaggle using Normalised Cumulative Discounted Gain (NDCG) [2]. The submission file to Kaggle contains a list of top 5 predicted destinations for every user in the test set.

7. Project Design

Project completion will be done in these steps:

- Making sense of the data provided. Visualizations to detect outliers, bogus data etc. All datasets will be studied and visualized for better understanding.
- Cleaning the data based on the previous step. Select datasets of importance and clean them. Either remove or impute unwanted columns, outliers if necessary.
- Merge datasets based on common columns and clean/impute if necessary.
- Split the training data into test set and training set for cross-validation.
- Consider different supervised ML models to work with and select the best option.
- Train a model based on the dataset and get a preliminary prediction using the selected ML algorithm and test it based on some basic metrics.
- Improve the model with cross-validation and GridSearchCV and test them with f1_score
- Once an acceptable score is obtained. Predict on the test data provided by Kaggle, create the submission file and upload to Kaggle for evaluating against the benchmark model.
- Repeat optimization and evaluation to improve the score.

Tools and Libraries used: PyCharm, Jupyter Notebook, pandas, sci-kit learn, seaborn [3], matplotlib. Other libraries will be added if necessary.

References

[1] Airbnb, "Airbnb About Us".
URL https://www.airbnb.ca/about/about-us

[2] Kaggle, "Kaggle : Airbnb Recruiting New User Bookings" (2017).
URL https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/details/evaluation

[3] M. Waskom, "seaborn" (2017).
URL http://seaborn.pydata.org