1. Introduction

We use Airbnb new user bookings dataset in Kaggle, trying to figure out what would be the first destination of based on information of Airbnb users. My main portion for the final project is data exploration and Naïve Bayes model.

2. Description (background information of algorithm)

Naïve Bayes

Naïve Bayes is a classification technique based on conditional probability, which often used as a baseline for more complex models. The basic assumptions of using Naïve Bayes are independence between features and all features contributing equally to the target. Most of other deep learning methods would treat features with difference importance. Let us take a closer look at the algorithm of this method.

Based on Bayes Rule, we could calculate the probability of class given certain condition of features.

$$\mathrm{P(A|B)} = \frac{P(B|A)P(A)}{P(B)}$$

$$\mathrm{P(Class|Features)} = \frac{P(\mathrm{Features}|\mathrm{Class})P(\mathrm{Class})}{P(\mathrm{Features})}$$

$P(\mathrm{Features}|\mathrm{Class}) - -$ conditional probability of the features given the class
$P(\mathrm{Class}) - -$ probability of the class
$P(\mathrm{Features}) - -$ probability of features

From previous data (train data), we could calculate the conditional probability of the features given the class label, probability of the class and probability of features. By independence assumption, probability of predictors could be estimated directly by multiplying the individual relative frequencies of each predictor

$$\mathrm{P(C|F)} = \frac{[P(\mathrm{f1}|\mathrm{C})P(\mathrm{f2}|\mathrm{C}) \ldots P(\mathrm{f1}|\mathrm{C})]P(\mathrm{Class})}{P(\mathrm{F})}$$

3. Description & 4. Results

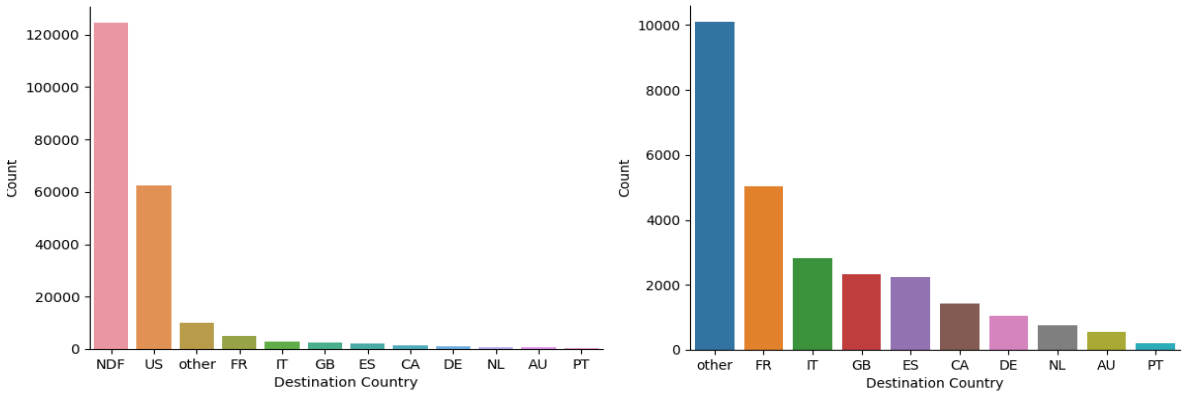
Data Exploration

Airbnb new user bookings dataset in Kaggle provides first booking destination decisions of users with their information. By exploring the data, we want to predict the first booking destination more accurately based on users information. In the train dataset, there are 213451 users with their decision of first booking destination and 15 features, which include 1. date variables: account created date, first active date, first booking date; 2. basic information: user id,

gender, age, language preference; 3. access to Airbnb: sign up method, sign up flow, affiliate channel, affiliate provider, first affiliate tracked, sign up app, first device type and first browser . In the test dataset, there are 62096 users with 15 features which are same as the train dataset, while first booking date are all missing value. Before creating features, let us take a closer look at the train data.

1). Destinations

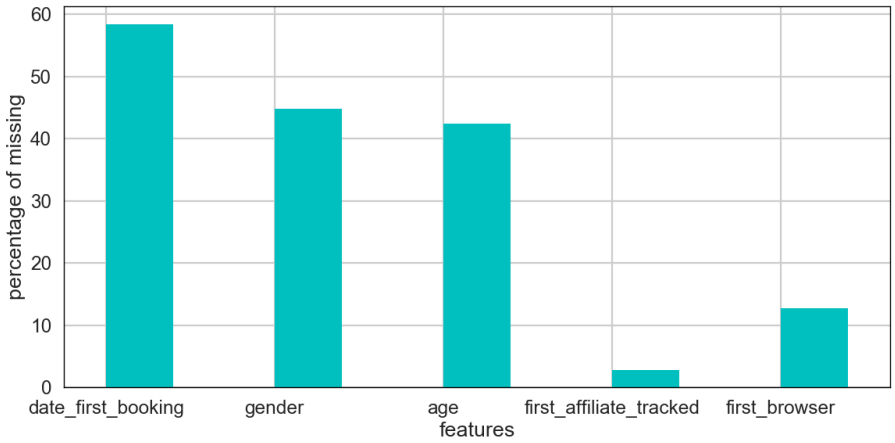
Among all 213451 users in the train data, 58.35% of users just browsing but have not booked in Airbnb yet; 29.22% of users chose the United States as their first destination. The remaining 12.43% of users chose many different countries as destinations, whose distribution is shown in below figure. The destinations' location and their counts are also shown in the map with pins in the location and marked with total population of who choose it as first destination. We also could know from the map that except the U.S., Canada and Australia, most of users in Airbnb would choose areas around Europe as their first destination.



2). Missing value

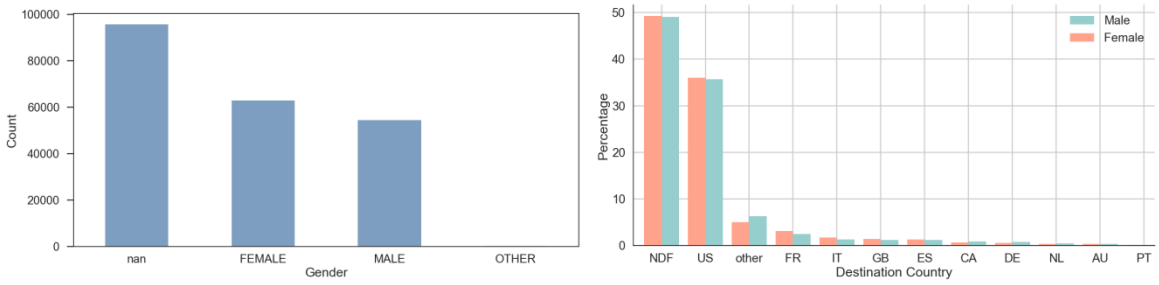
After replace "unknown" as missing value, the percentages of missing value

in features with missing value are shown in figure below. As the missing value percentage of first booking date in train data is nearly 60% and there is no first date booking information in test data, we will delete this feature in model training.



3). Basic information

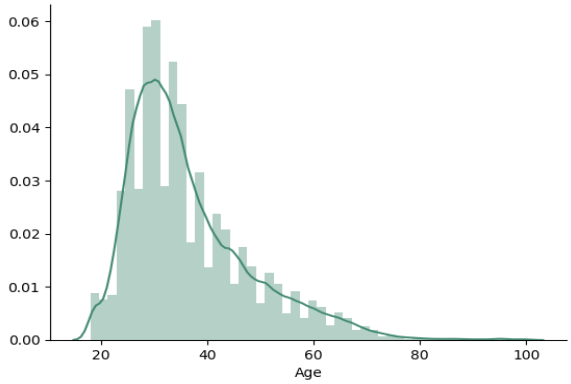
First, we look at the distribution of gender. It seems like many people do not want to specify their gender in user information. As the number of Nan (which replace unknown) is larger than male and female, we would take it as a separate category in gender feature. Among those who specify their gender, female users are larger than male users. And when we look at the gender distribution on destination, we find that male and female do not have big difference in the first destination decision.

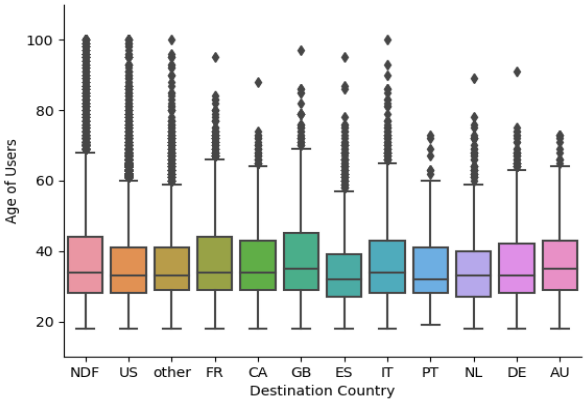


By calculating we know 96.66% of users in the train dataset prefer to use English. That might be why except for those who did not book, most of people chose the United States as their first destination.

Age value shows that there might be some input error of this feature, because Airbnb only allow people who are 18 or older to create an account, but there are 158 users below age 18. Also, the max of age is 2014, which is unreasonable as an age number. So we guess some values might be recorded wrongfully with birth year instead of age. After dealing with those outliers, the distribution of age between 18 and 100 shows that users of Airbnb centers on age 25 to age 40. When we look at the age distribution in different country destinations, Spain seems to

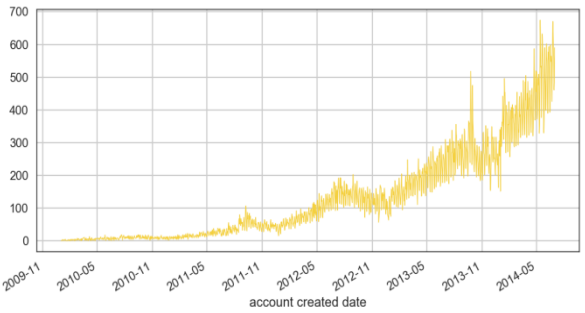
be chosen by more younger users as their first destination in Airbnb, the Great Britain seems attract users in a wider range of age.

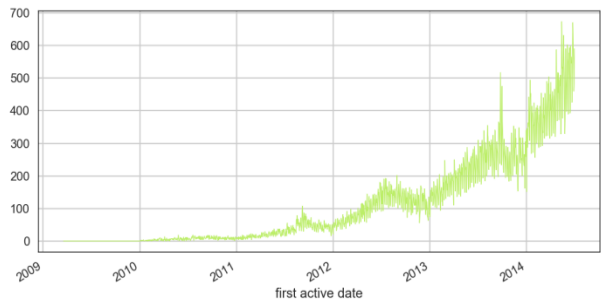




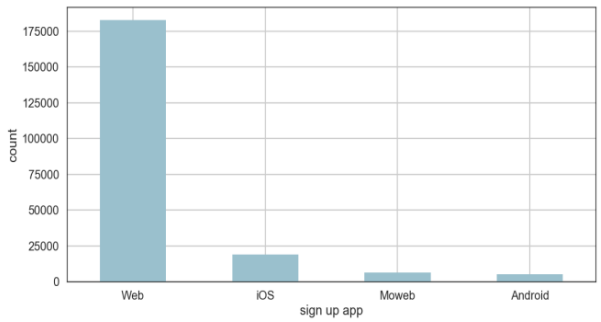
4). Date features

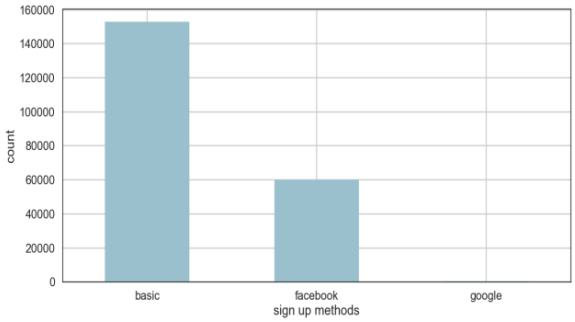
Except for feature -- first booking date, there are two features include date information: account created date and first active date. First active date could be earlier than account created date, as people might browse before signing up. After plotting them out, we find that the distributions of these two features are pretty similar and both of them have obviously increasing trend. Both of these two plots indicate that Airbnb thrive after year 2012. The wave in two figure also indicates that August to October seem to be most active period among each year.

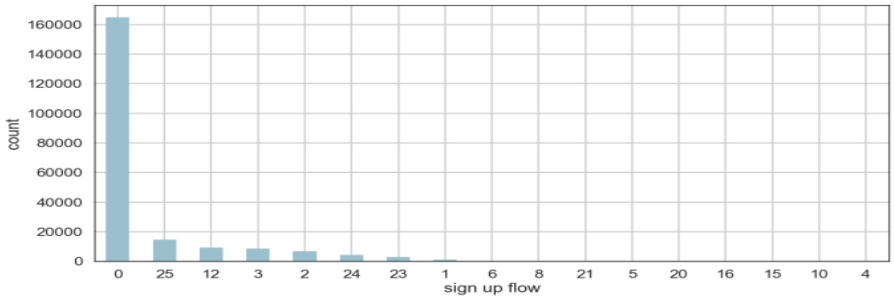




5) Other features

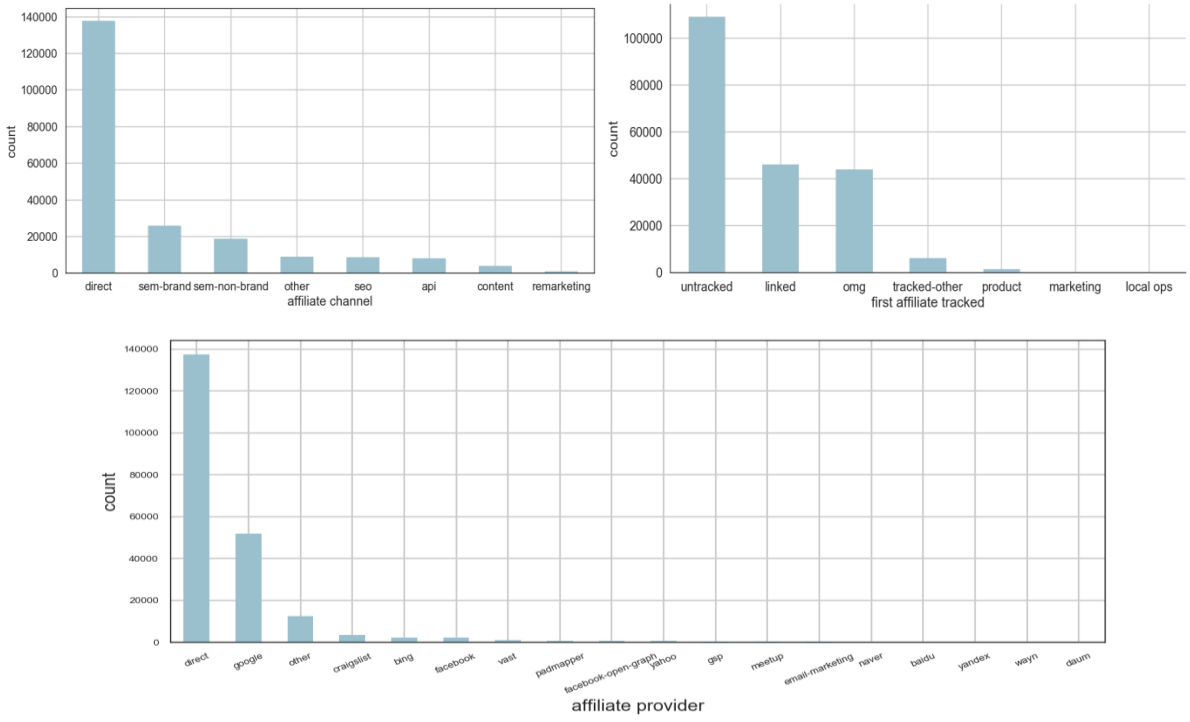




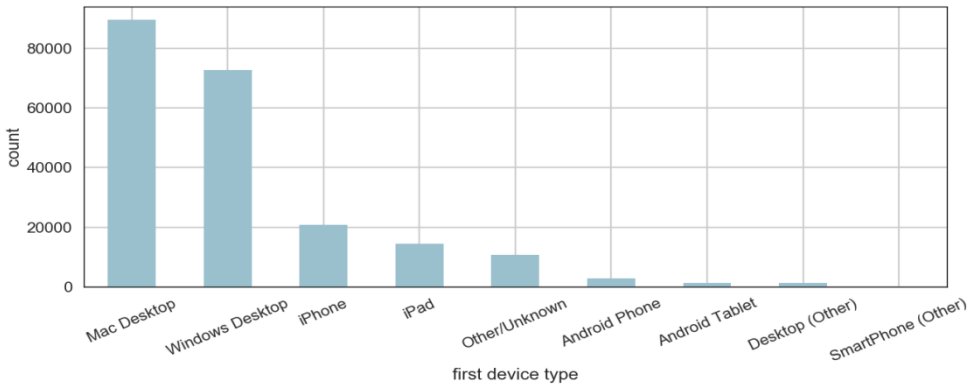


The first three figures above represent the distributions of three features about Airbnb account sign up: sign up app, sign up method and sign up flow. By calculation, we know that 85.6% of users signed up through web, the other three accesses – iOS, Moweb and Android only share with a small percentage. For sign up method, 71.63% signed up by basic method, 28.11% signed up by Facebook and only 0.26% signed up by google. For the sign up flow, we know that 77.18% of users came to signup up from the page 0.

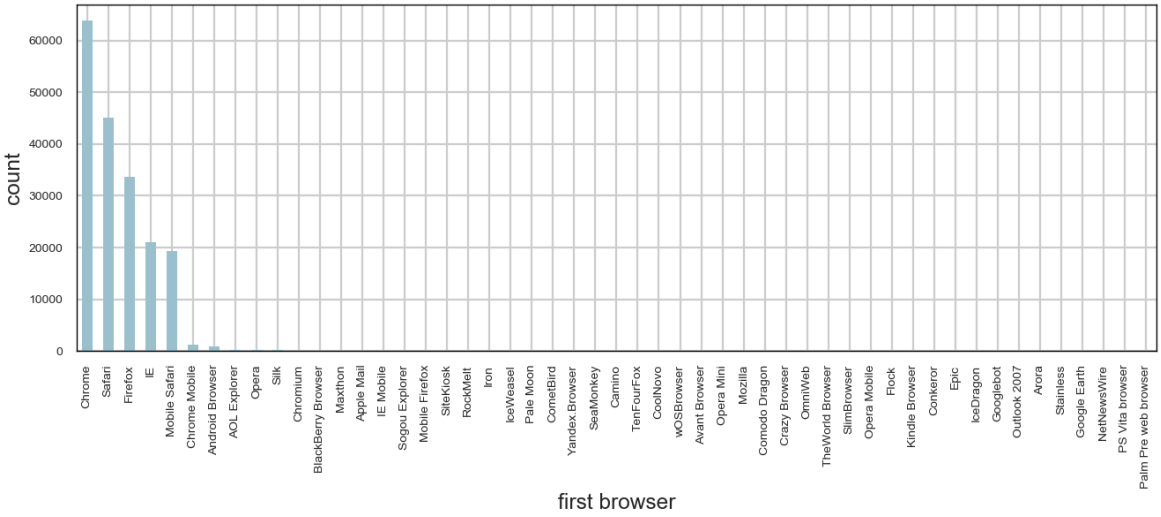
Three figures below show the distribution of three features about affiliate. By calculating their percentage in each category, we find most users came to Airbnb directly instead of going through other affiliates. There is 64.52% of users are marked as "direct" in affiliate channel; 64.38% of users are marked as "direct" in affiliate provider.



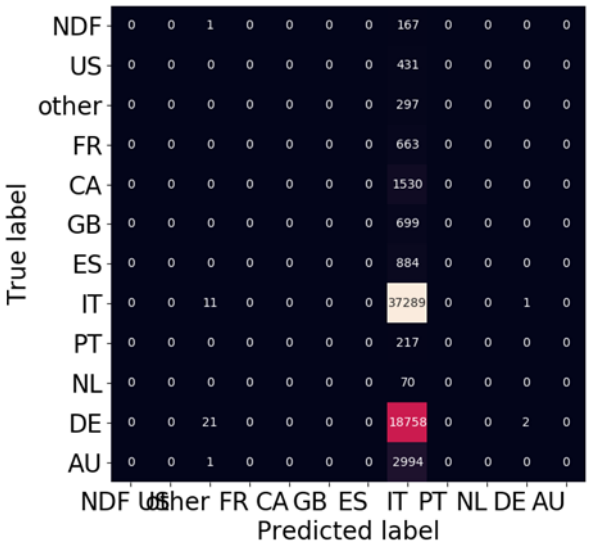
From the distribution of first device type feature, we know Mac Desktop users account for 41.98%, which is the highest share. Besides, Windows Desktop users account for 34.07%, iPhone account for 9.73%, iPad accounts for 6.72% among whole users.



From the distribution of users' first browser, we know 29.91% of users using Chrome to access Airbnb, 21.16% using Safari, which are two main shares among all the browsers listed in the train data. Also, Firefox accounts for 15.77%, IE accounts for 9.87% and Mobile Safari accounts for 9.03%.



Result of Naïve Bayes



We split data into 30% and 70% two groups. The group has 70% of whole train data as train, 30% of train data as X test. After fitting the Naïve Bayes, we get an

accuracy of 58.25 and confusion matrix of prediction on X _test, which shown in above. The accuracy on X test is pretty low. We guest it is because some features in the data are not independence with each other, and some features might not have a stationary mean or standard deviation. From the confusion matrix, we could find that when predicting for label, only Italy and Germany have correct true positive prediction. Also, most of true label Germany ones would be wrongfully labels as Italy.

Also, after upload the file into Kaggle, we get a score of 0.84255. It is a little bit weird that we get a pretty low accuracy in X test but getting a high score in Kaggle.

Name	Submitted	Wait time	Execution time	Score
sub_nb1.csv	5 hours ago	0 seconds	9 seconds	0.84255

5. Summary and conclusion

Exploring data is a long and hard way, although we could learn by others' experience, when we want to explain them and visualize them, we need to try many different ways to take a closer look at it. By exploring the data, I have learn to set many customizes that I did not use before to help me plot figures with better appearance.

6. Percentage

60%

7. Reference

[1] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). New York: springer.