Booking-Go

CMPN451: Big Data and Analytics

Submitted To: Eng. Ahmed Youssry

Team (9)

Group Members

- Ahmed Hany
- Areej Khalid
- Mariam Amr
- Nadin Magdy

Table of Contents

Introduction & problem description

Recently, online businesses have been focused on enhancing user experiences and attracting the right audience by utilizing the data that is provided by social media websites, cookies, and other means. If instagram businesses can do this, we can apply it to travel businesses. Airbnb is the most famous website for booking places to stay with more than half a billion guests per year.

This means that not only do we have big data produced from the user base, but also this user base can be used to predict where a person will travel by their age, gender, and browsing habits.

In this project, we have records of information about 12 countries, and around 210k+ users of Airbnb all from America. With their information, we can predict where a person is going to travel first. This means we will be able to keep up with the world making personalized user experiences for Airbnb new customers, capturing users' attention by the ease of booking and promoting the places they are more likely to travel to, thus creating a good reputation and increasing the user base of the business.

Project pipeline

First of all, we start with data preprocessing including dropping irrelevant columns like ID, we have a column name called timestamps we extract from it season, months, day, day-week new features, replacing Nan values with realistic values.

We also used Hadoop MapReduce integrated with google collab where Map gives (Key,Value) pairs of (Season feature, the rest of features) which was then sent to the reducer to perform some of data cleaning in age feature where age feature has the highest data cleaning weight, we divide the ages into groups and handle irrelevant values there. The output from the reducer was exported as a .csv file to be sent to model.ipynb to continue data encoding and normalization and finally perform models training and compare their results.

Analysis and Solutions

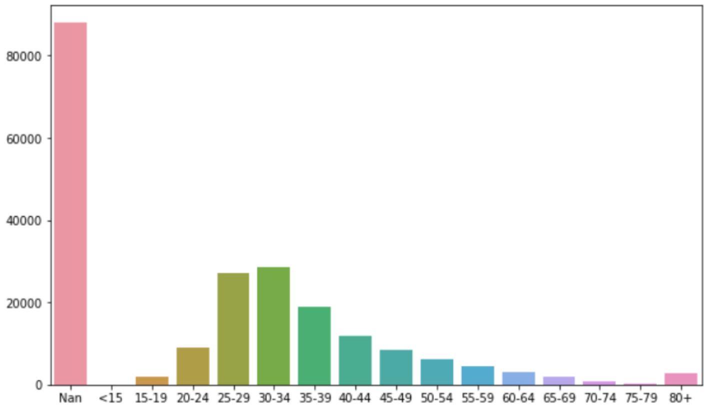
1. Data description and preprocessing

Our data from the Kaggle data set contains several files for the purpose of this prediction:

- Age gender buckets: shows the population in thousands of people of a specific age range, living in a specific country, with a gender X, at a year Y
- Countries which as our 12 countries we classify by, their geographical position, the language they speak, etc.
- Sessions
- User data is the main data used for the model. It has the user's Ids in Airbnb, and information like gender, age, account creation data, first booking date, sign-up browser, the language they used in the browser, etc. Our data has more than 210k rows.

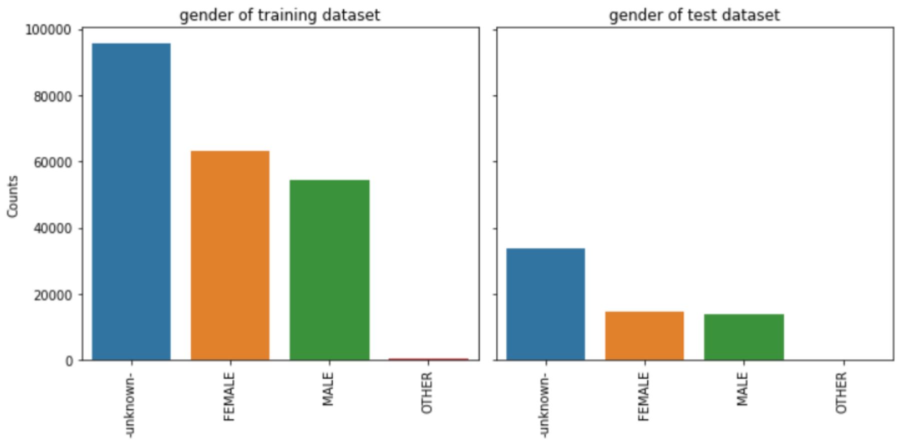
2. Data visualization

First, we split our users from the user data into age buckets we use to visualize the ages we have. Almost half of our data is unknown, and otherwise, we see a normal distribution in the ages, where between 25-35 are the majority of users' ages.

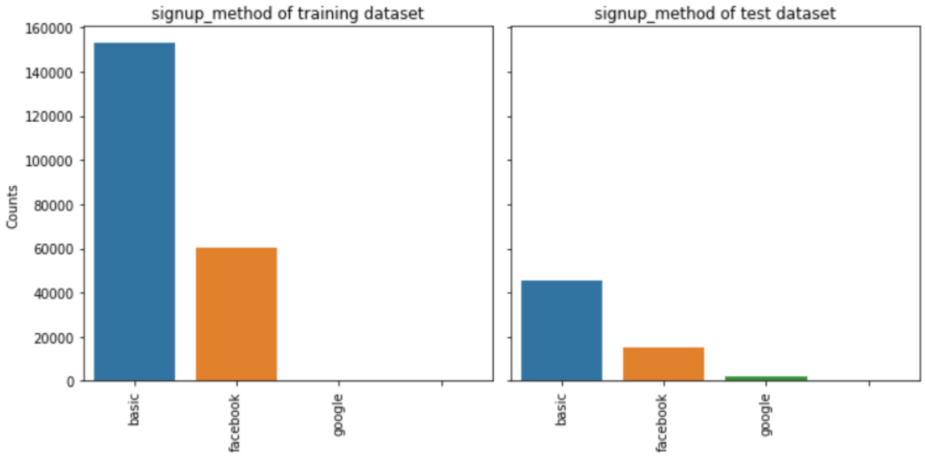


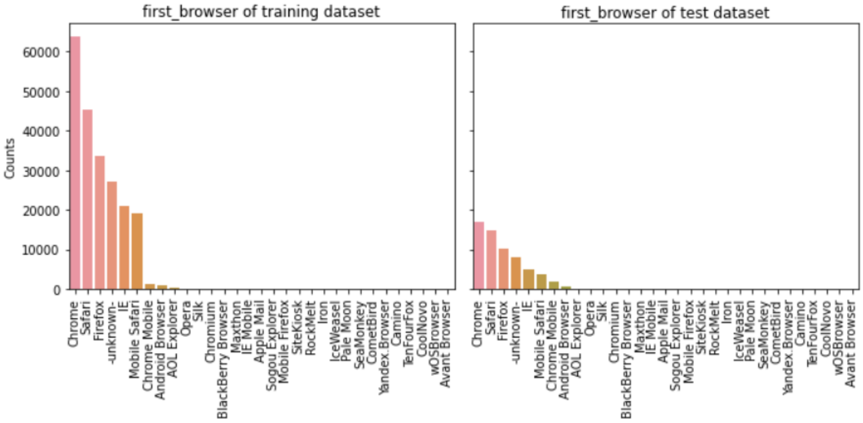
We also have the gender feature, which is also mostly unknown. We had several options when dealing with it:

- Replacing unknown gender with a probability of frequency of each gender, so if there are 60% female, random set 60% of unknown to female
- Because the unknowns are so many and we can see as a user behavior in itself, we can choose to keep it as it is. We chose to go with this option

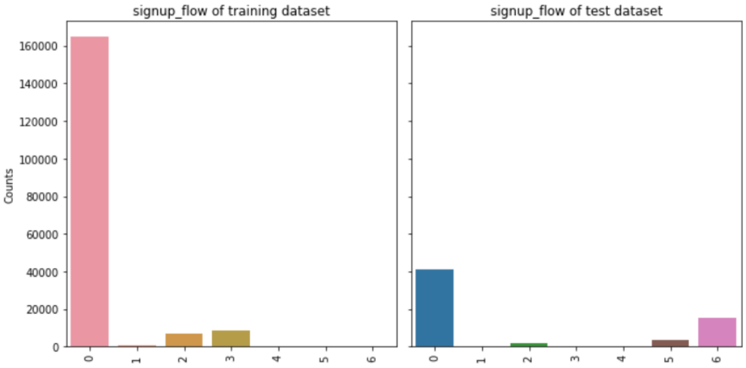


Here we have a problem with the value ‘google’. It is almost nonexistent in the training set but we find it a lot in the test set. The model will not know enough information about this value and the results for it might affect the model, if this feature is importan**t.**



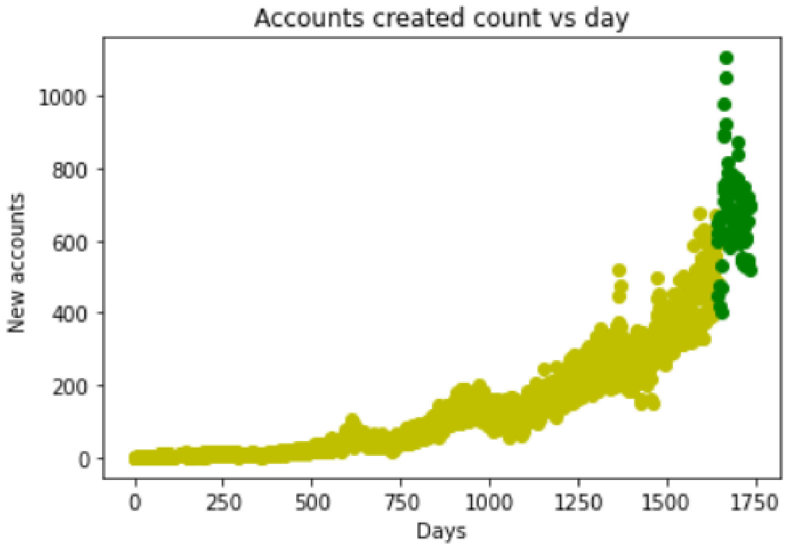


Same problem for 2 and 5 here as the 'google' problem.



We see almost see exponential growth from the day of the first account being made (day 0) up to the latest accounts being made.

The yellow are the accounts that were made till a certain day, and the green are the new user accounts made recently and we are trying to personalize the booking experience for.



3. **Extracting insights from data**

When we first tried to predict the country we looked at the correlation between all the variables and the country destination. We found some of the most important variables to be**:**

- The first active timestamp: time of first activity reflects on the seriousness of booking, and what country is in season for visiting
- Sign up method: for instance, signing up from desktop makes you more likely to book a certain destination
- Language: using Airbnb in a certain language means you are more comfortable with this language, and you may be looking to book in a country where they speak this langauge
- Sign up flow
- The month of the first booking

4. **Model/Classifier**

- In order to find a good multi-class predictor for this particular dataset,

we applied 5-fold cross-validation using GridSearch against 3 classifiers (AdaBoot, Multi-linear perceptron, QuadraticDiscriminantAnalysis) with various parameter configurations. After the evaluation is done, we report the best scores and then we train the classifiers with the best parameters.

We tried prediction using several models:

- KNeighborsClassifier with 58% accuracy
- Decision tree with 33%
- Random forest with 64%
- Gradient boosting with 64%
- Adaboost with 64% confidence
- Linear Discriminant Analysis with 62%
- Quadratic Discriminant Analysis with 80%
- Multi Linear Perceptron with 81%

Conclusion

The best model to predict the user booking is Quadratic Discriminant Analysis

We have an accuracy in prediction of 80%. We decided to choose QDA as it is much faster than MLP with almost the same final accuracy

References

[1] https://www.kaggle.com/competitions/airbnb-recruiting-new-user-bookings/data