Capstone Project Jyothsna Krishnamurthy

Machine Learning Nanodegree 02-09-2018

Domain Background

This opportunity is taken from the Kaggle project

https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings

The project is used to predict the destination(upto 5) a new user will make his next trip based on web session records provided.

Airbnb was founded in August 2008 and is based in San Francisco , California. The company's primary focus is to enable customers to book accommodations online or using mobile app from accommodations offered by other hosts. These accommodations can be someone's second home, extra living space etc. Airbnb offers unique accommodations that fits each customers price point. Airbnb has an international presence in about 65,000 cities and 191 countries.

A brief synopsis about the data science team in Airbnb:- Airbnb started a small data science team (with 1 person) in 2010. Initially the data science team was perceived as a team that mostly provided statistics like, How many listings we have in Paris? Or What are the top 10 destinations in Italy? While these numbers are important Airbnb recognized the importance of the sequence of events that leads up to getting those numbers. In their own words "A datum is a record of an action or event, which in most cases reflects a decision made by a person. If you can recreate the sequence of events leading up to that decision, you can learn from it; it's an indirect way of the person telling you what they like and don't like — this property is more attractive than that one, I find these features useful but those — not so much." Hence it was not just important to understand the numbers but the voices of the customers behind those numbers was a way to perceive the data.

Airbnb believes in a more proactive partnership between decision makes and the data science team. This enables a very closely knit cross collaboration between the data science team and the decision makers. Adding one more impressive line from their published article. Our distinction between good and great is impact — using insights to influence decisions and ensuring that the decisions had the intended effect. While this may seem obvious, it doesn't happen naturally — when data scientists are pressed for time, they have a tendency to toss the results of an analysis "over the wall" and then move on to the next problem.

From then on Airbnb has done a lot in the field of data science to enable them to be successful and focus more on data driven decision making. Below is a link to an article from where some of the quotes highlighted have been taken, it a great read!

https://venturebeat.com/2015/06/30/how-we-scaled-data-science-to-all-sides-of-airbnb-over-5-years-of-hypergrowth/

The datasets used for this project are provided by Kaggle and Airbnb – link below

Link : https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data

Below is a link to a white paper which uses machine learning implemented via random forest to solve a multi class – classification problem on classifying the type of disease

https://www.sciencedirect.com/science/article/pii/S2214317316300099

Problem Statement

This problem was picked from a Kaggle competition from the Kaggle website. The problem is to accurately predict where a new user will book their first booking. Airbnb has grown so much in the last few years because it has connected so many hosts to many customers looking for customized accommodations in specific prize ranges for their visit to a particular location whether the reason for travelling is vacation, business or something else. This unique experience has opened some new vistas in the travelling experience very different from the hotel options available in the past. It has also provided a competitive market and also allows hosts to make money from the facilities they have available.

My personal motivation for this problem is my passion for data , I love the fact that the company does not treat data as mere statistics or numbers. Data will tell you a story if you look into it carefully. From the article link in the domain background Airbnb has been true to this statement. Just like in everything they do, in this problem they have provided information about the web sessions of the users leading upto the decision point of choosing a particular destination. With the information provided in the sessions file I can train a model find patterns and use that trained model to accurately forecast where the customers destination will be. I look forward to using the data exploration skills and modeling skills acquired from my study in Udacity to see how I fair. I am a sql person and data analysis is not new but the interesting combination here is my newly acquired python skills along with the newly acquired modeling skills /methodology (data science modeling) to complete this project makes it challenging and interesting to me. It will amaze me at the end to see how I started out with just some data sets and tuning/predicting goals to be able to use some techniques to arrive at an optimal model.

The problem will be able to clearly make 5 predictions (1st one is the most probable one ordered by the remaining 4 less probable) of the location of the next destination of the user using what has been learned from the web sessions. From the problem statement the user can be predicted to go to one of the 12 destination countries 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF' (no destination found). The evaluation metric provided is NDCG (normalized discounted cumulative gain) @ k. Where k is the number of a prediction , 5 predictions are expected k takes a value from 1 thru 5. The country the person is predicted to choose first gets a NDCG = 1. The order of the remaining 4 countries and the NDCG value should be much less. The evaluation metric is discussed in detail in the evaluation section of this project proposal.

Datasets and Input

The following data sets are provided in Kaggle for this project

Link : https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data

The data set consists of a list of users , their demographics and web session records, and some summary statistics. All the users from the dataset are from the USA.

The training and the test data sets are split by dates.

- train_users.csv - the training set of users
- test_users.csv - the test set of users
  - id: user id
  - date_account_created: the date of account creation
  - timestamp_first_active: timestamp of the first activity, note that it can be earlier than date_account_created or date_first_booking because a user can search before signing up
  - date_first_booking: date of first booking
  - gender
  - age
  - signup_method
  - signup_flow: the page a user came to signup up from
  - language: international language preference
  - affiliate_channel: what kind of paid marketing
  - affiliate_provider: where the marketing is e.g. google, craigslist, other
  - first_affiliate_tracked: whats the first marketing the user interacted with before the signing up
  - signup_app
  - first_device_type
  - first_browser
  - country_destination: this is the target variable you are to predict
- sessions.csv - web sessions log for users
  - user_id: to be joined with the column 'id' in users table
  - action
  - action_type
  - action_detail
  - device_type
  - secs_elapsed
  -

- countries.csv - summary statistics of destination countries in this dataset and their locations
- age_gender_bkts.csv - summary statistics of users' age group, gender, country of destination
- sample_submission.csv - correct format for submitting predictions.

Based on the initial data analysis the number of records in given for training are - 213451,
The classes distributed based on the target variable are as below

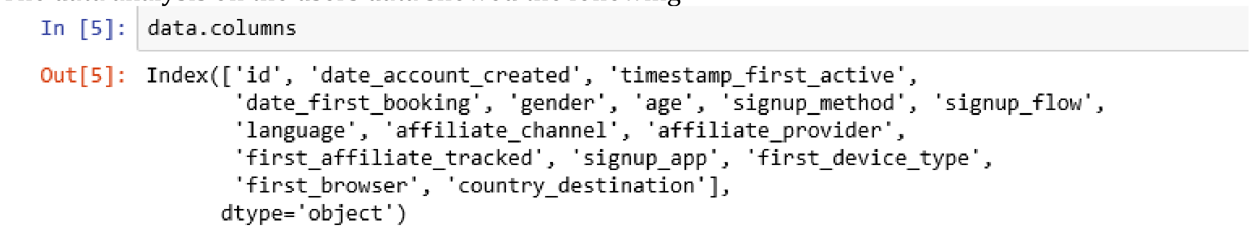
```
country_destination
```

```
NDF      124543
US        62376
other     10094
FR         5023
IT         2835
GB         2324
ES         2249
CA         1428
DE         1061
NL          762
AU          539
PT          217
```

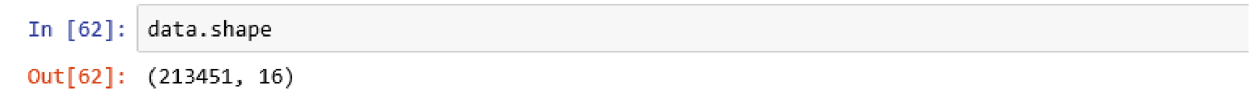
The thought process is to split the data using Stratified Kfold to preserve the classes. Stratified Kfold splits the data into folds by preserving the percentage of classes in the fold, 1 fold is used as set each time.

Data Exploration

The data analysis on the users data showed the following



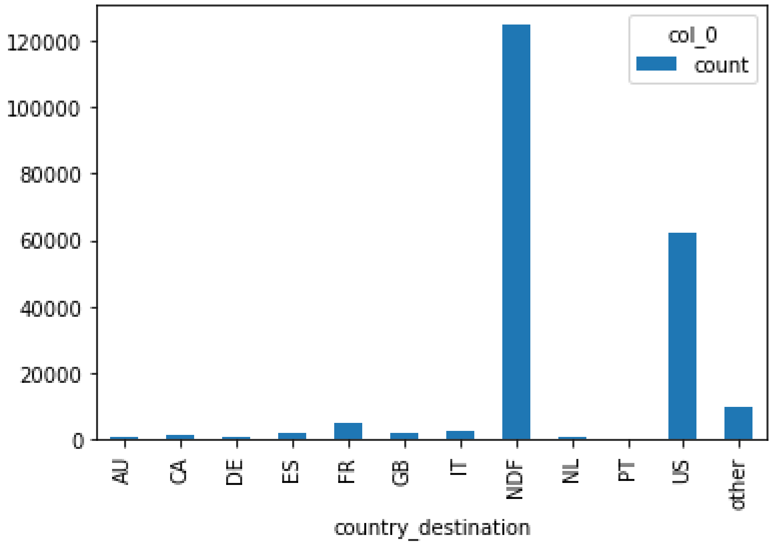
```
In [5]:  data.columns

Out[5]:  Index(['id', 'date_account_created', 'timestamp_first_active',
                'date_first_booking', 'gender', 'age', 'signup_method', 'signup_flow',
                'language', 'affiliate_channel', 'affiliate_provider',
                'first_affiliate_tracked', 'signup_app', 'first_device_type',
                'first_browser', 'country_destination'],
               dtype='object')
```



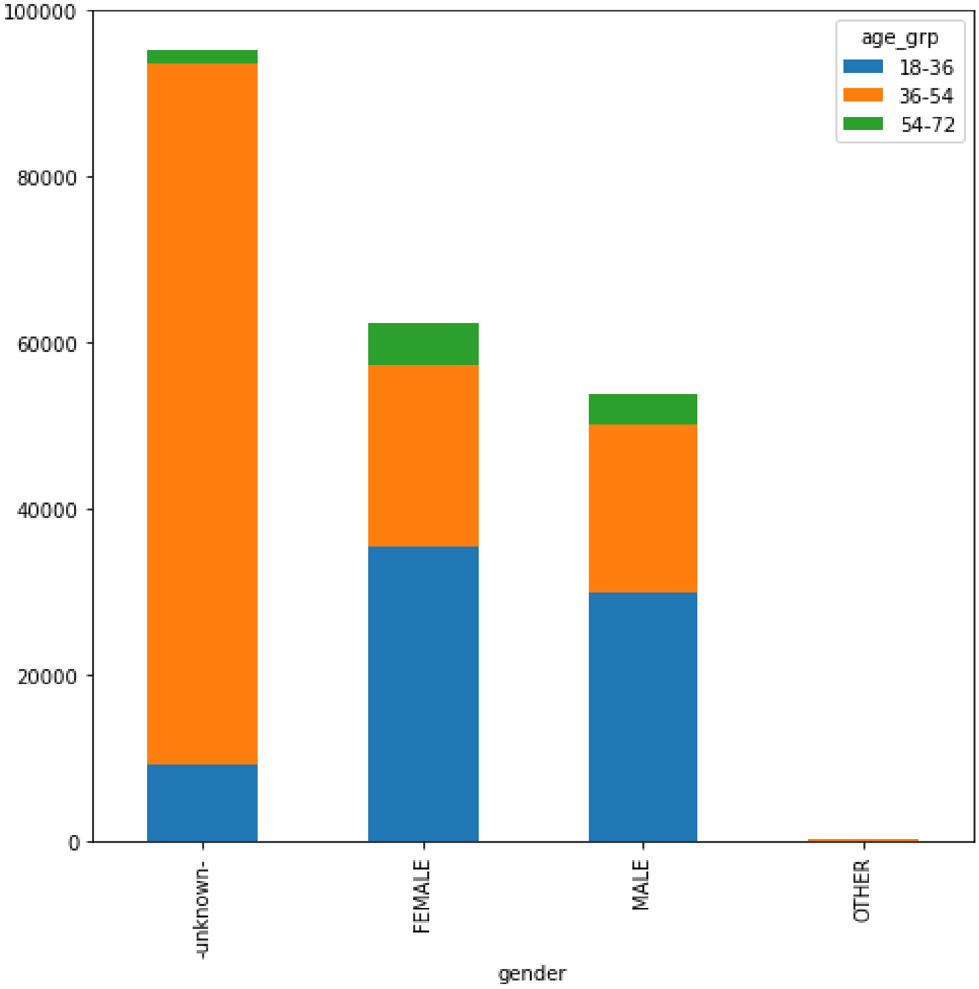
```
In [62]:  data.shape

Out[62]:  (213451, 16)
```

The users data had about 213451 unique user ids.

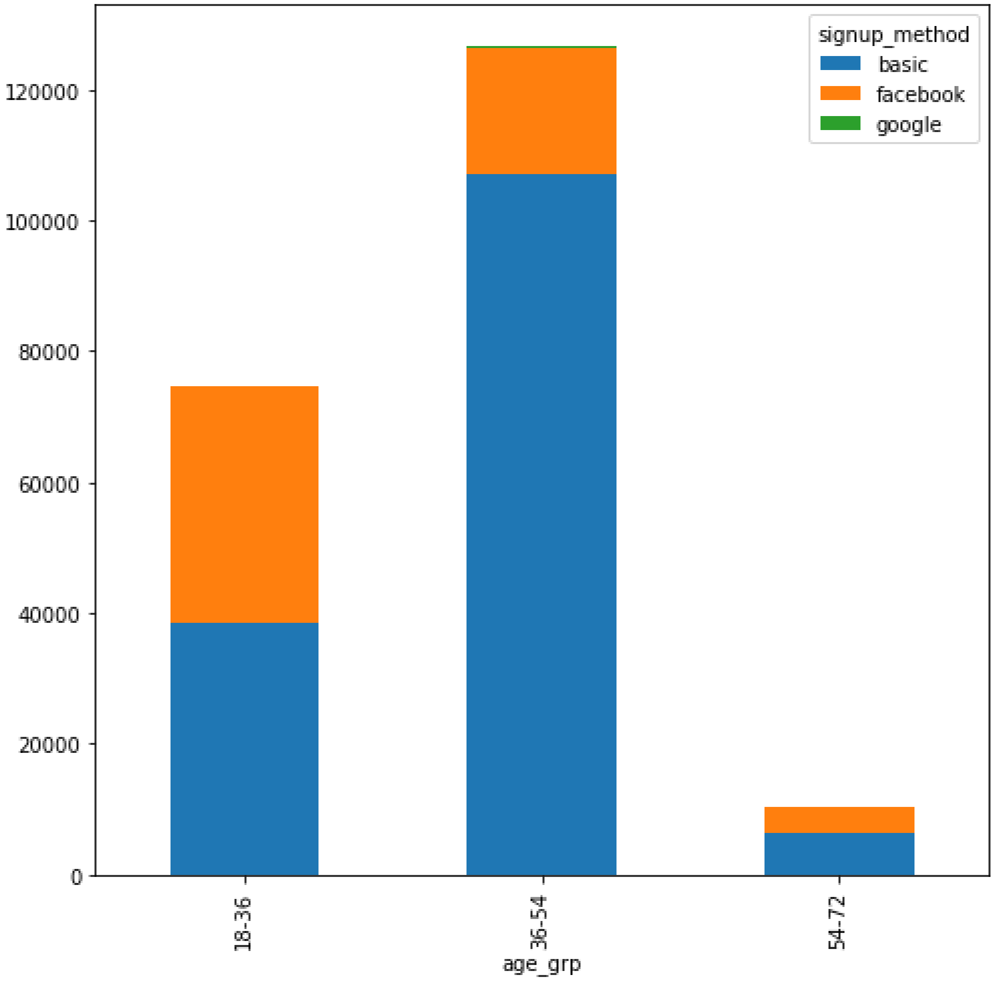
Exploring the **target variable country of destination** it can be seen that after the 'NDF' , US is the next highest place travelled followed by 'Other'. It also seems like Paris , Great Britain and Italy are also places of destination that AirBnb travelers chose.



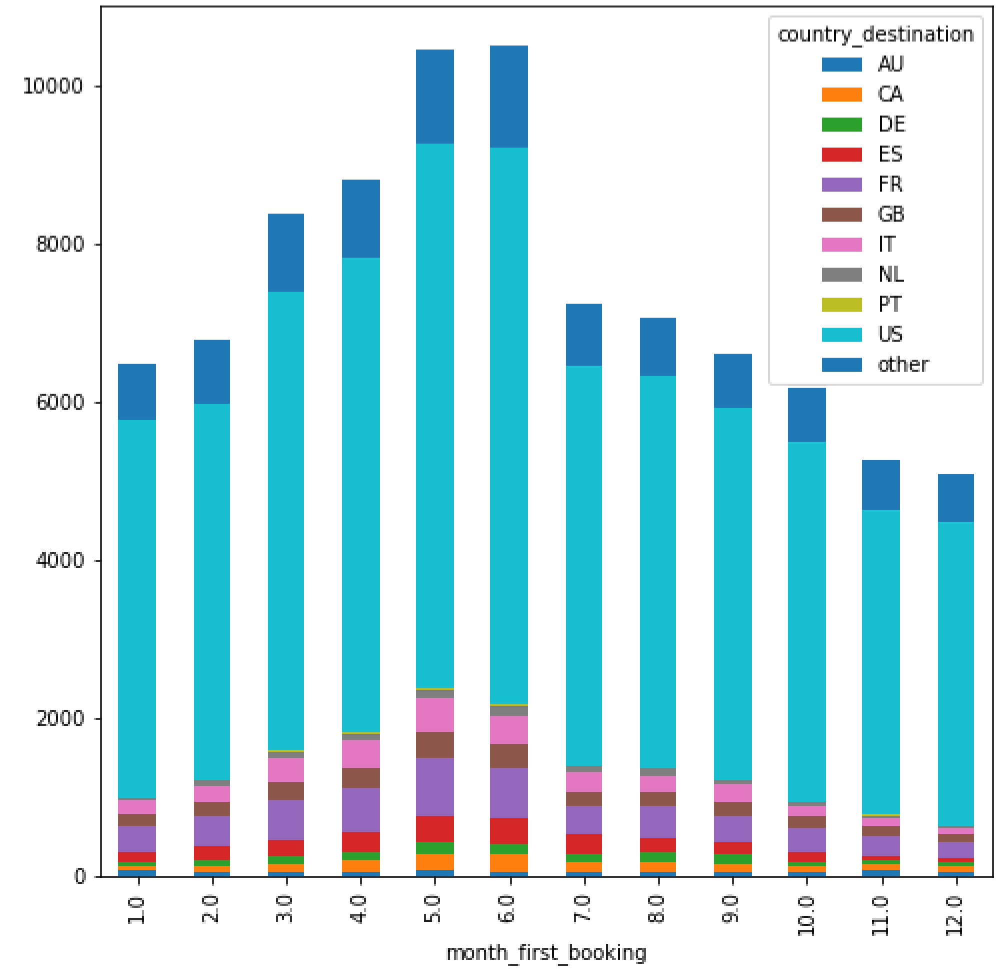
Gender age group analysis , shows that there are more people in the age group 36 -54 . It can be seen that in both female and male popultion there are less people in the age group 54 -72



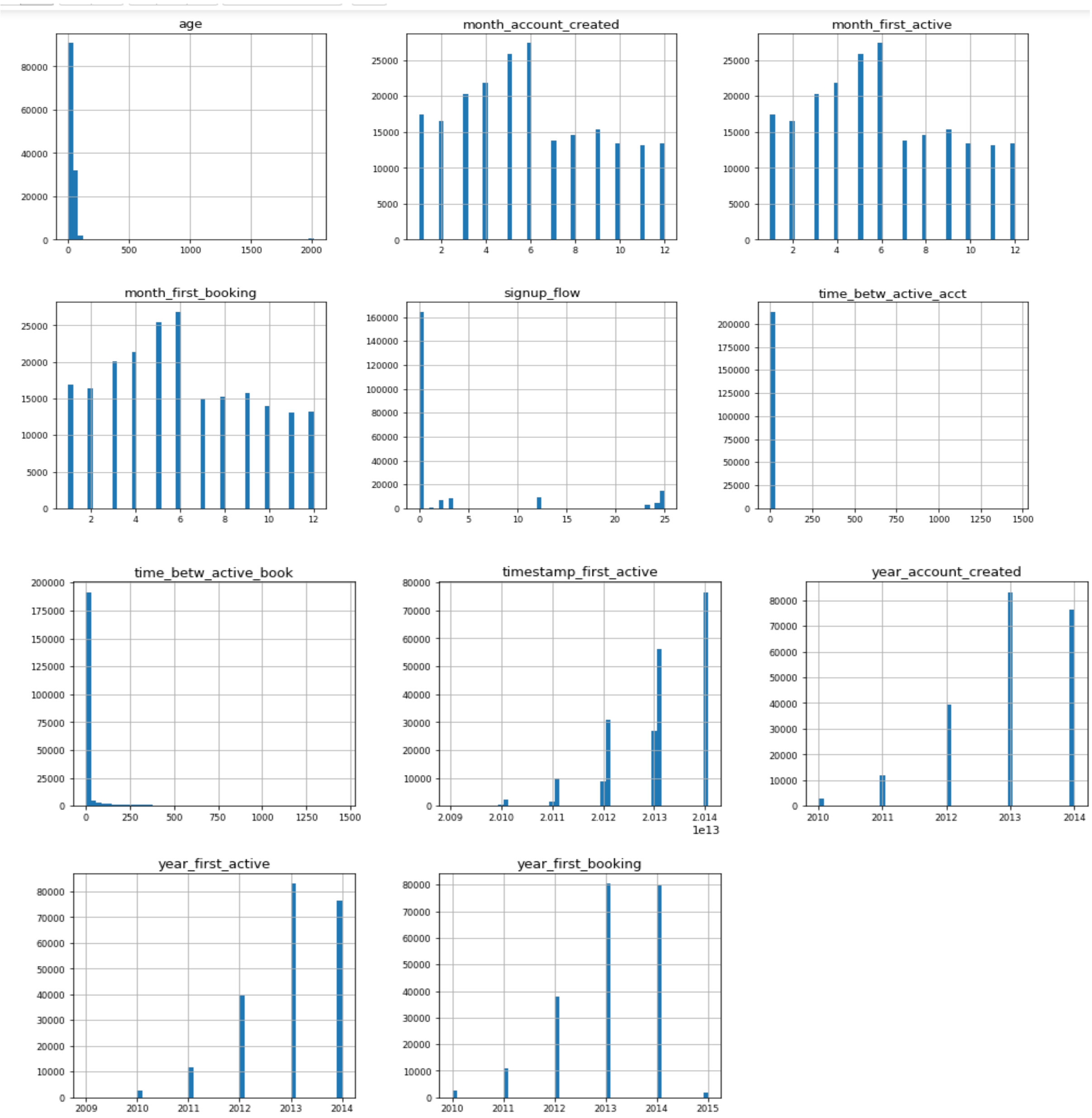
In the exploration of the of the gender and the sign up method it can be seen that google is rarely used as the signup method. The “unknown” group mostly uses basic while the “male” and “female” groups use both facebook and basic almost equally.
Comparing sign_up method and age it can be seen that google is the least preferred method. Most people in the age group of 36-54 prefer basic method of sign up. Age group 18 -36 use facebook and basic methods for sign up. Least signup is noticed in the age group 54-72



It can be seen that high number of bookings happen in the 5th and 6th month of the year , May and June. The booking increase from Jan to May peaks in June and then decreases from June to December. December is the holiday season so people are busy with family and friends so booking numbers are at the lowest. It can also be seen in the below graph that US is the highest chosen destination.

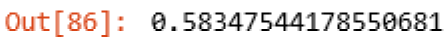


The below graphs are analysis of numerical values even though they are categorical in nature provide some good insights. It can be seen that the months of account creation, first active and first booking follow a similar pattern peaking mid year and trailing off towards the end of the year. It can be seen that more accounts were created in the year 2013 but 2014 was the most active year. It can also be seen that 2013 and 2014 were good years when users made the most bookings.



Benchmark Model

To Benchmark the results , the model was treated for removal of Nans . The model was then trained used random forest. The random forest provided the below accuracy score

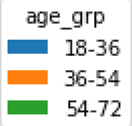


```
Out[86]: 0.58347544178550681
```

Data Preprocessing step

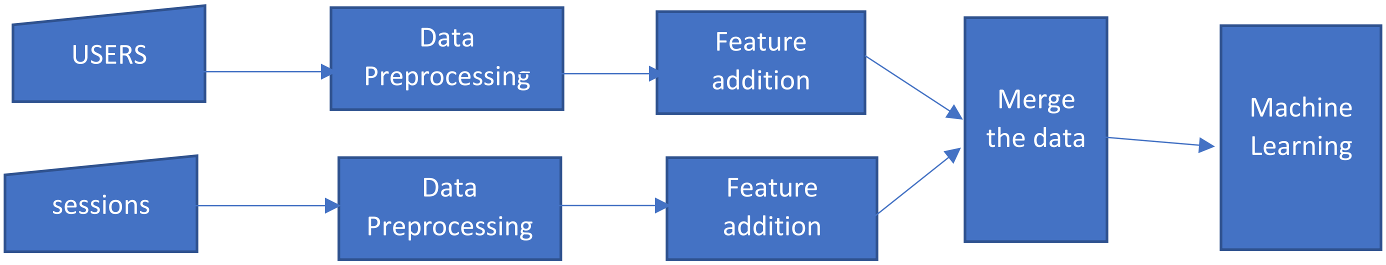
The preprocessing that needed to take place before the data could be trained were

- Treat the Nulls in one of the date columns
- Split the data columns into month and year
- The age column had Nulls and outliers
  - The null values in the age column were imputed by the mean of the age
  - The outliers were also substituted by Nan which were imputed by the mean of the age
- The age column had a large number of values hence it was binned using pd.cut into age brackets to make it more categorical -

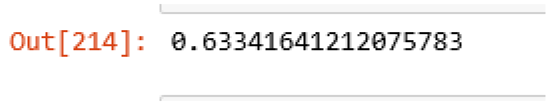
- The differences between the timestamps added more features to the data. The days between creating an account and booking, days between first active and booking. days between first active and booking.
- The days obtained from the differences between the timestamps were further binned using pd.cut into '<3m', '3 -6m','6-9m' ,'9m-1y', '>1y']. The upper and lower boundaries is dependent on the max and min value of the dataframe column
- The sessions data was first treated for null values.
- The secs elapsed in the session data was converted to hours
- The hours were then grouped into ranges
- The session data was then merged with the primary user data by user id

Implementation



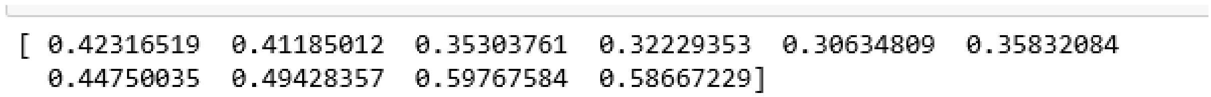
The users data was treated for basic null values . This data was then trained on a Random forest model and a benchmark score was obtained.
The data was then pre processed and feature engineered using the steps mentioned in the pre processing.
Initially from the sessions table the action, action detail column were used to pivot and summarize the data, but due to the large number of columns and high data volume the device type column was used
The device type was used to summarize the sessions data
The sessions data was eventually merged with the users data using the USER ID column.
This data was fed into the machine learning model KNN and Randomforest classifier to get the score.
The KNN and RandomForest with the preprocessed and featured data had a score of

Random Forest



```
Out[214]: 0.63341641212075783
```

KNN using cross validation with cv = 10



```
[ 0.42316519  0.41185012  0.35303761  0.32229353  0.30634809  0.35832084
  0.44750035  0.49428357  0.59767584  0.58667229]
```

Results

It can be seen from the benchmark and the models that the scoring is higher in both cases than the benchmark model.

Model	Score
Benchmark – random forest	0.58347544178550681
Tuned – random forest	0.63341641212075783
KNN best score	0.59767584

From the above it can be seen that the tuned model yields better results than the benchmark model.

Conclusion

The model has been trained using 213451 rows of user data and sessions data. The model has the max of 63% accuracy. The necessary pre-processing steps have been applied to ensure that a good amount of data is available for training and generalizes well.
There are some refinements that could be done to improve the model accuracy, the sessions data has columns like action and action detail which when pivoted have a lot of distinct values , action has about 360 distinct values. Due to the limited capacity of the machine I am working not all features could be engineered as needed.
Also the data provided also had a lot of Unknowns, Null values and the target variable had a lot of NDF (not defined). Providing better quality data will further improve the model

References

https://venturebeat.com/2015/06/30/how-we-scaled-data-science-to-all-sides-of-airbnb-over-5-years-of-hypergrowth/

https://machinelearningmastery.com/handle-missing-data-python/

https://www.kaggle.com/<userid>/data-sciencetutorial-for-beginners/editnb

https://www.dataquest.io/blog/pandas-cheat-sheet/