**Group #232: Airbnb Data Analysis**

|  |  |  |  |
| --- | --- | --- | --- |
| **First Name** | **Last Name** | **IIT email** | **Student ID** |
| Akhil | Padgilwar | apadgilwar@hawk.iit.edu | A20427219 |
| Nupur | Singh | nsingh26@hawk.iit.edu | A20431624 |
| Lian | Liu | lliu81@hawk.iit.edu | A20415798 |

Table of Contents

[1. Introduction 4](#_Toc8053371)

[2. Data Sets 4](#_Toc8053372)

[3. Problems to be solved 5](#_Toc8053373)

[4. Solutions 5](#_Toc8053374)

[5. Exploratory Analysis 6](#_Toc8053375)

[5.1. Age Analysis 6](#_Toc8053381)

[5.2. Users Gender Distribution 6](#_Toc8053382)

[5.3. Users Language Distribution 7](#_Toc8053383)

[5.4. Using sign up method distribution 7](#_Toc8053384)

[5.5. Signup app distribution 8](#_Toc8053385)

[5.6. Label (histogram and Boxplot) 8](#_Toc8053386)

[6. Experiments and Result 9](#_Toc8053387)

[6.1. Load Data into Python 9](#_Toc8053389)

[6.2. Missing values chart 9](#_Toc8053390)

[6.3. The spread of Country Destination 10](#_Toc8053391)

[6.4. Data preparation: 10](#_Toc8053392)

[6.5. Model Build 15](#_Toc8053404)

[7. Conclusion 21](#_Toc8053412)

[8. Future Work 21](#_Toc8053413)

# Introduction

As we all know, Airbnb is an app that helps people book house for traveling. People on Airbnb can book a place to stay in 34,000+ cities across 190+ countries. No matter where you go, you are able to use Airbnb to find a house. Millions of travelers choose Airbnb as a platform to plan their travel. It is very popular all over the world. Furthermore, more and more new users start to use Airbnb to build their trip.

This project is based on the analysis done by Airbnb where they collected the data of first time users on the basis of their demographic features and other details to understand which country did they choose as their first destination.

In order to build models, we make some preprocesses to our data first. Then, we build some classification models based on prepared data. We use accuracy as metrics to evaluate our model. We found the best model with the highest accuracy.

# Data Sets

Airbnb data set is the collection of user’s information of over 210000 users between Jan 2010 to Jun 2014. All these users are selected randomly. This data set has a total 16 attributes. The data set has some missing values; we need to perform preprocessing before we build a model.

|  |  |  |
| --- | --- | --- |
| Columns in the dataset | Data Type | Explanation |
| ID | Categorical | User’s unique ID |
| date\_account\_created | Numeric | The date of account creation |
| timestamp\_first\_active | Numeric | Timestamp of the first activity |
| date\_first\_booking | Numeric | Date of first booking |
| Gender | Categorical | User’s Gender |
| Age | Numeric | User’s Age |
| Signup\_method | Categorical | The way to sign up |
| Signup\_flow | Numeric | The page a user came to sign up from |
| Language | Categorical | International language preference |
| affiliate\_channel | Categorical | what kind of paid marketing |
| affiliate\_provider | Categorical | Where the marketing is e.g. google, craigslist, other |
| first\_affiliate\_tracked | Categorical | What is the first marketing the user interacted with before the signing up |
| Signup\_app | Categorical | What kind of app to first sign up |
| First\_device\_type | Categorical | What device people use to sign up first |
| First\_browser | Categorical | What kind of browser people use to sign up first |
| Country\_destination | Categorical | The final destination people choose. |

The label for this data set is Country destination. There are 12 possible outcomes of the destination country: 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF' (no destination found), and 'other'. 'NDF' is different from 'other' because 'other' means there was a booking, but to a country which is not included in the list, while 'NDF' means there wasn't a booking.

# Problems to be solved

We have built multiple classification algorithms on this data so that we are able to make a prediction for a new user’s destination on Airbnb. This can help Airbnb give a more efficient recommendation to new users.

# Solutions

Data Preparation: For our original data, there are too many missing values. We have counted how many missing values in each feature. And we use NA to represent them rather than null. We have transformed category data to numerical data for the model.

Multiple algorithms: We build different models with different accuracy. We compare their accuracy to find the best model. The various model includes: Decision Tree, Naïve Bayes, Random Forest, SVM, K-Nearest Neighbor, and Ensemble Classifier.

# Exploratory Analysis

In this part, we build some graphs to show some analysis.



## Age Analysis

A close up of a map

Description automatically generated

From the picture, we can see that most users stay 20 to 40 this range. It means most users are young people.

## Users Gender Distribution

A screenshot of a cell phone

Description automatically generated

We can see that the number of females is more than the number of males. And we have some other users in another gender.

## Users Language Distribution

A picture containing screenshot

Description automatically generated

As the picture shows, most of the people use English as the preferred language. We can simply make a prediction that most of the people would like to choose the country in which the first language is English as their first destination.

## Using sign up method distribution

A screenshot of a cell phone

Description automatically generated

From the picture, we can see that over 10000 people choose basic to sign up to Airbnb which means they have an account in Airbnb. Almost 4000 people choose Facebook as their account to sign up to the Airbnb. Not more than 500 people choose google account to sign up Airbnb.

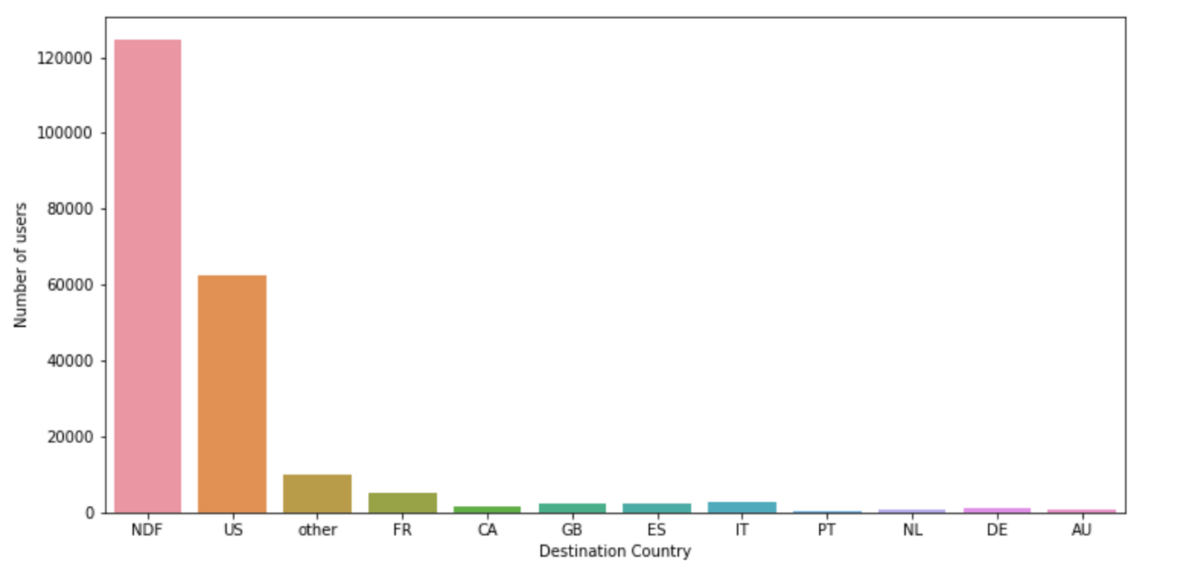
## Signup app distribution

A screenshot of a cell phone

Description automatically generated

As the picture shows, almost 12000 people choose the web to sign up to Airbnb which means they are highly possible to use a computer to browse the website. IOS is the second group that people would like to sign up to Airbnb. Moweb and Android is the third group to sign up to Airbnb.

## Label (histogram and Boxplot)

****

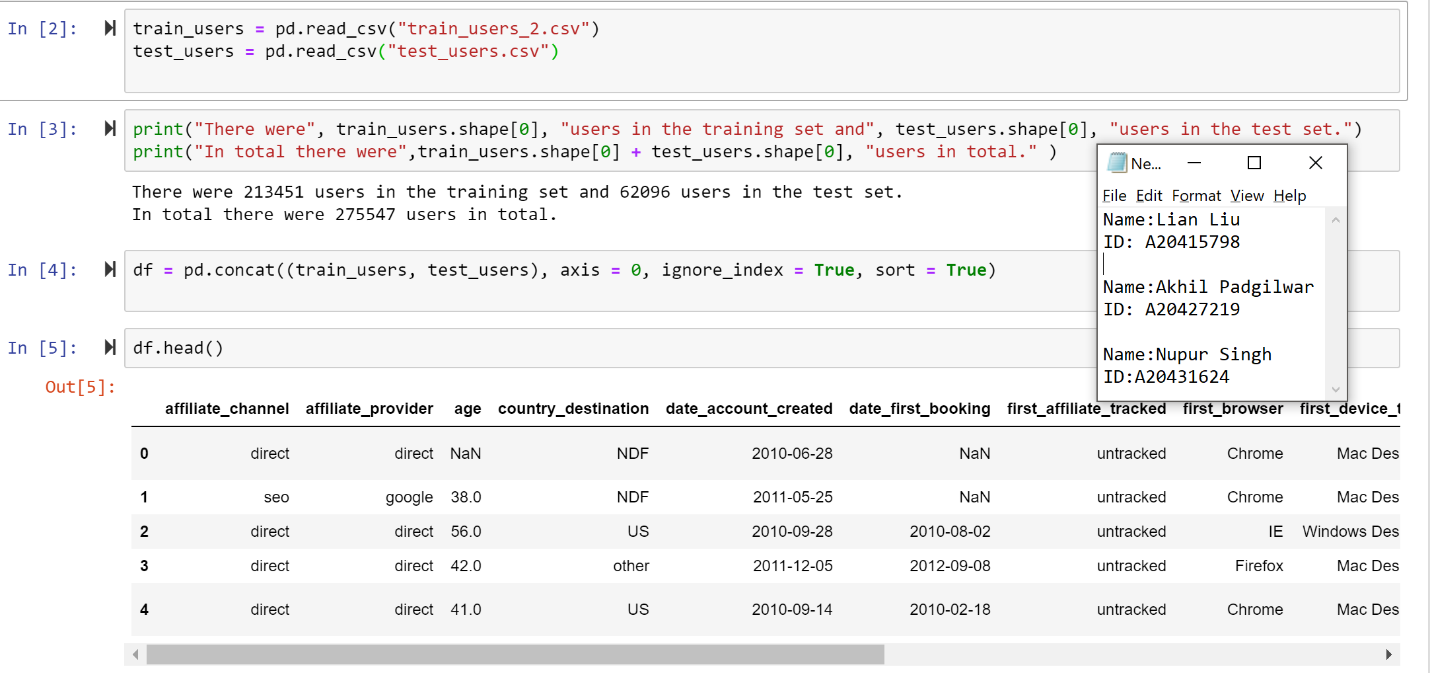
This shows that Nearly 60% of the users did not end up booking any trip represented by NDF. A majority of users booked a destination in US (about 30%) considering that user population in this problem is from US. Thus, my inference is that US travelers tend to travel within US itself.

# Experiments and Result



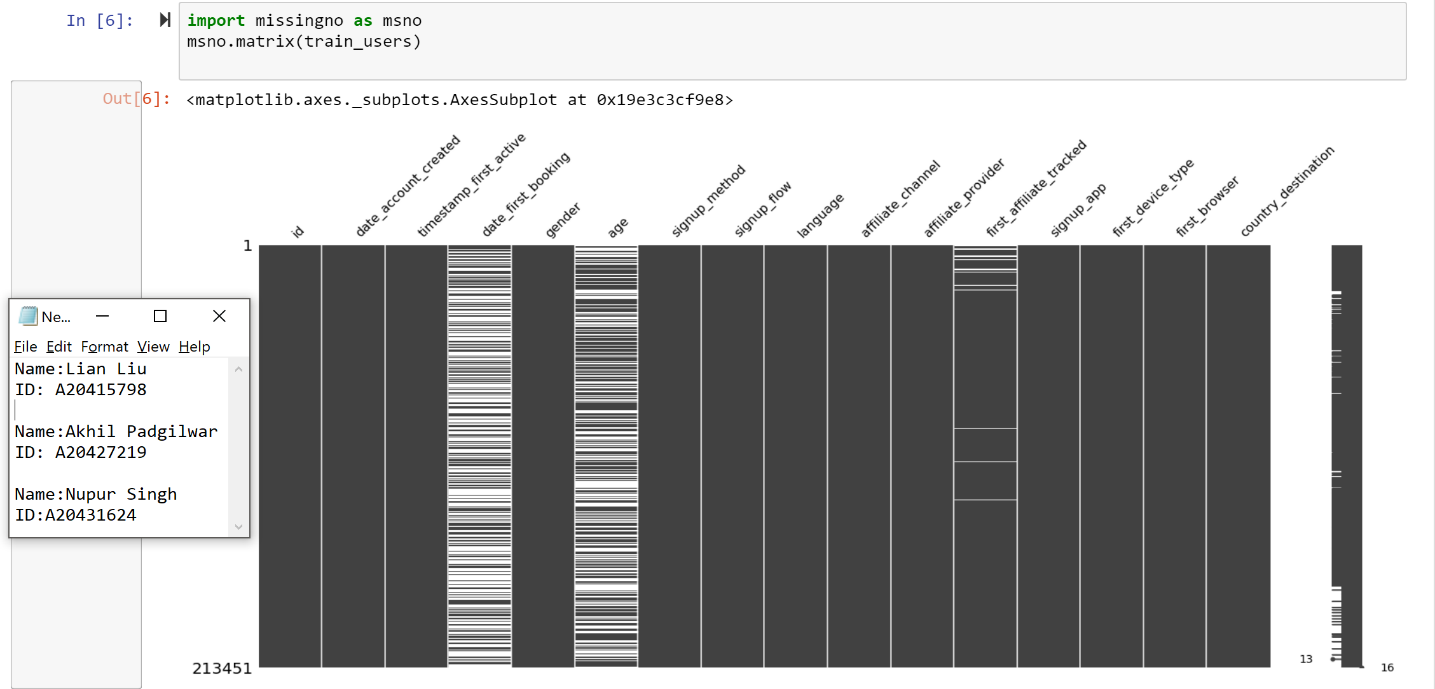
## Load Data into Python

We first load our data into Python and check how many rows data we have. And we use head() function to show the first 5 rows data.



## Missing values chart

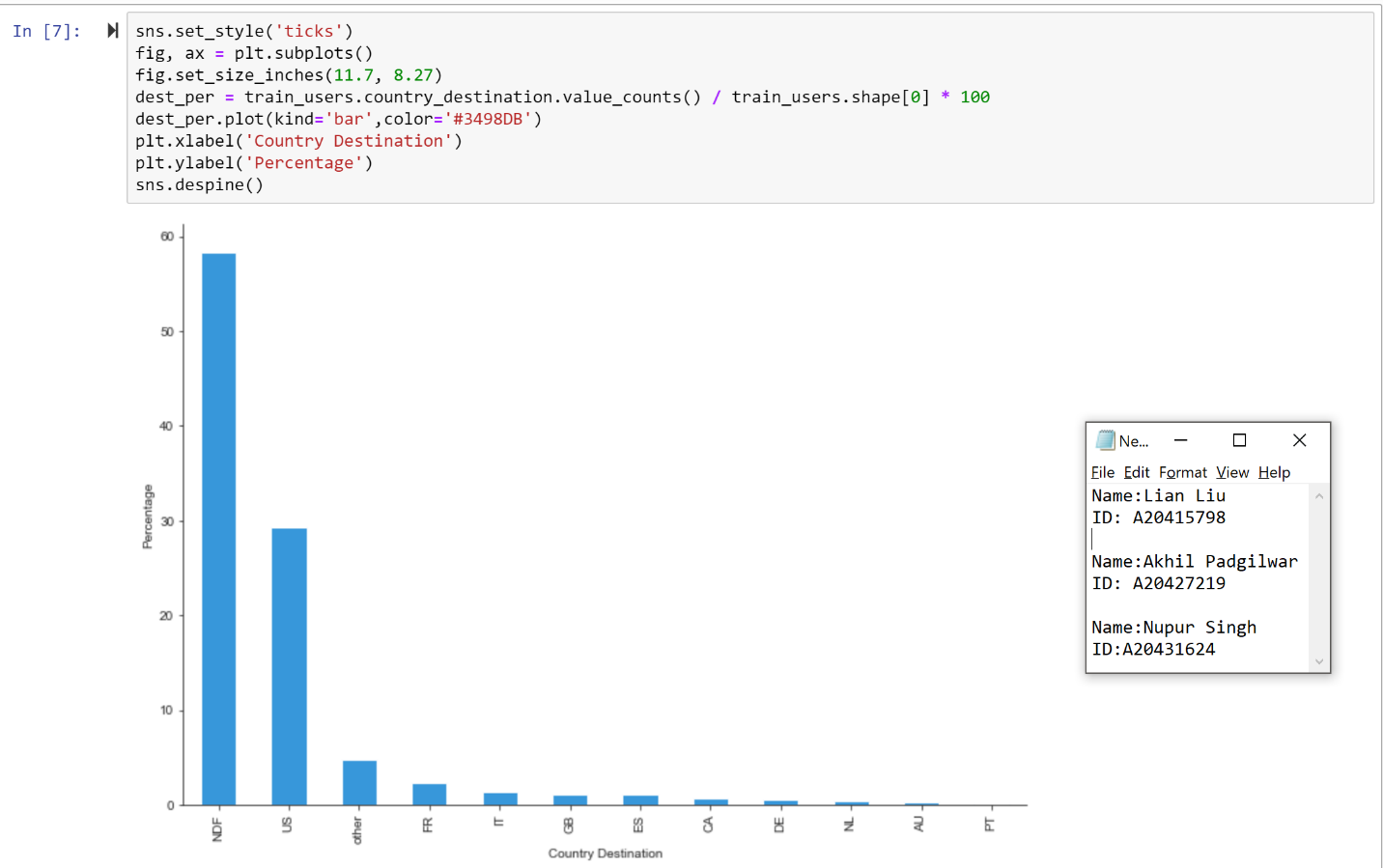
For a better model, we need to know how many missing values we have in each feature. So we build this chart to show missing values.



The gray part means no missing values. The white part means missing values. From the picture, we can see that date\_first\_booking and age have many missing values. First\_affiliate\_tracked have some missing values.

## The spread of Country Destination

We build a histogram to show the percentage of country destination.



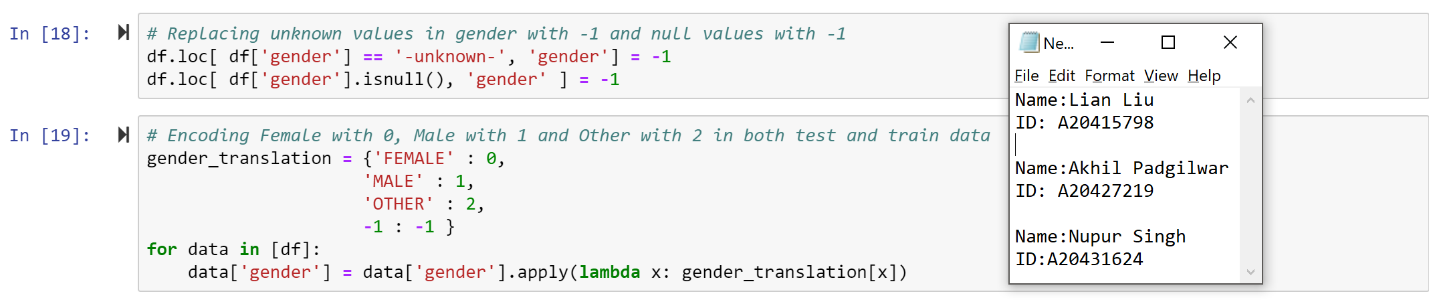
From the picture, we are able to know that NDF has the highest percentage. NDF means no destination found. It is almost 60% of the whole data. The second highest country is the US.

## Data preparation:

In our data, there is so many categorical variables. We need to transform them into a numerical variable. Because some models require all the features should be numerical variables.

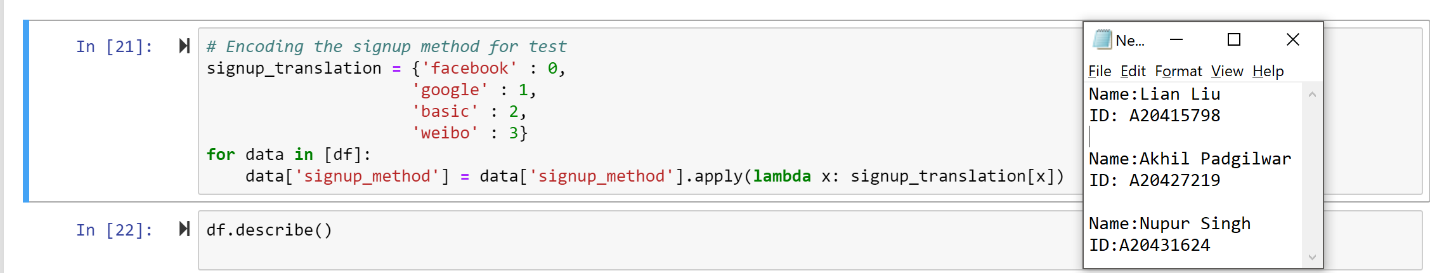
### Gender feature

From the dataset, we find that there is a value called unknown in gender feature. We cannot just remove those value or replace them with the most used value in that column. Because unknown means people do not choose their gender when they register the first time. Hence we kept it as it is in the data and decided to use -1 to represent unknown, 1 to represent male, 0 represent female and 2 represent to others.



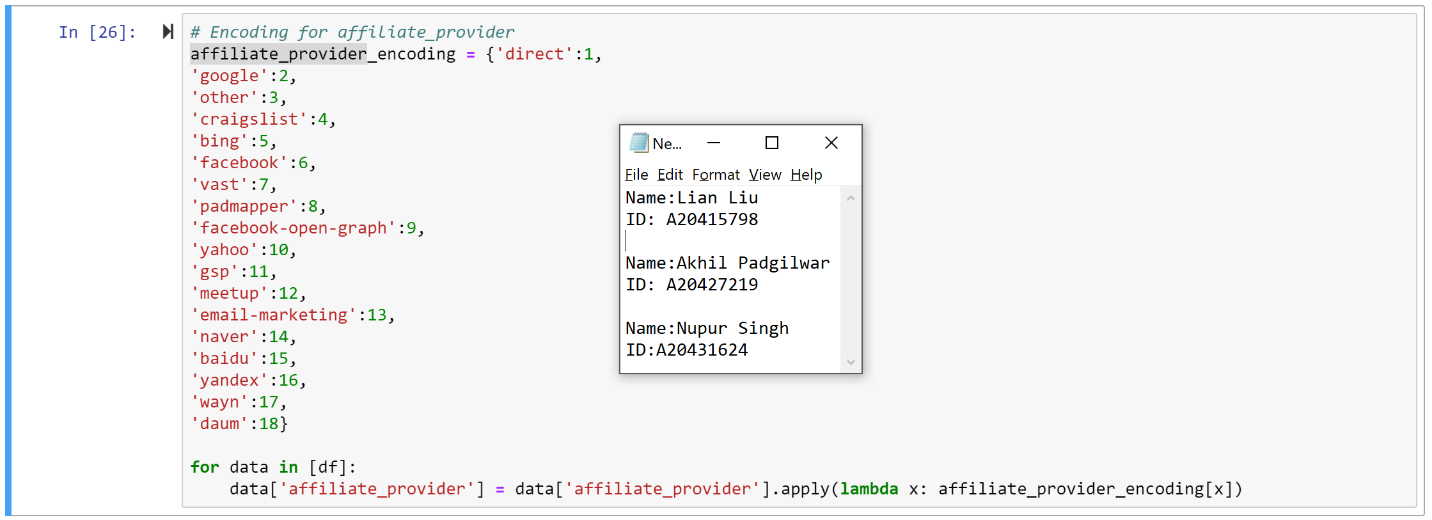
### Sign up feature

For sign up the column, there are 4 unique values, they are facebook, google, basic and weibo. We use 0-4 to represent them.



### Affiliate\_provider feature

For this categorical feature, there are 18 unique values. We determine to use 1 to 18 to represent them.



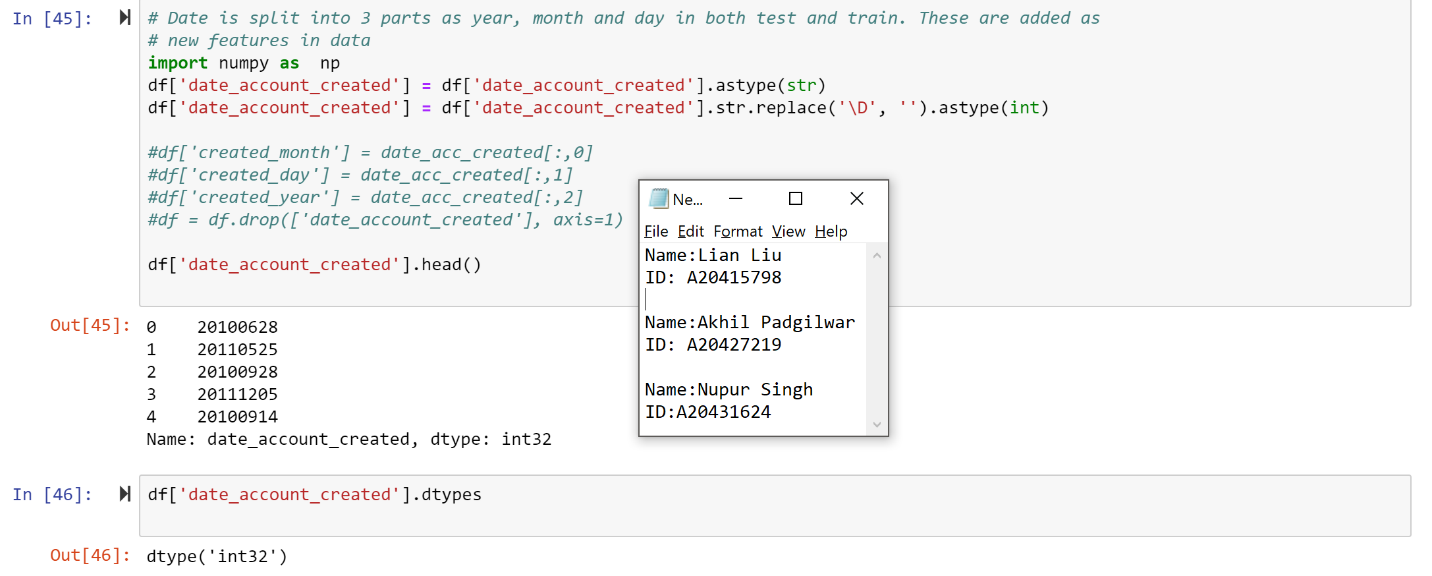
### First\_browser feature

From datasets, we are able to know that this feature has 56 unique features. So, we decide to use 1 to 56 to represent these values.



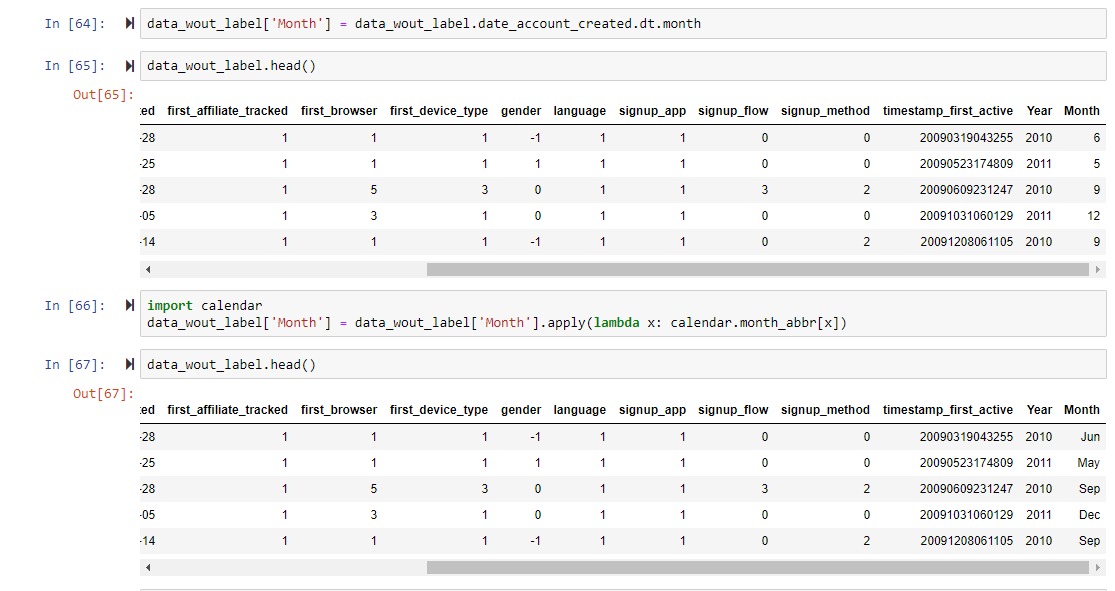
### Date\_account\_created 1

For this feature, the date is a time feature. However, python does not recognize these number as a numerical variable. They are recognized as time type. We need to transform them into numerical data.

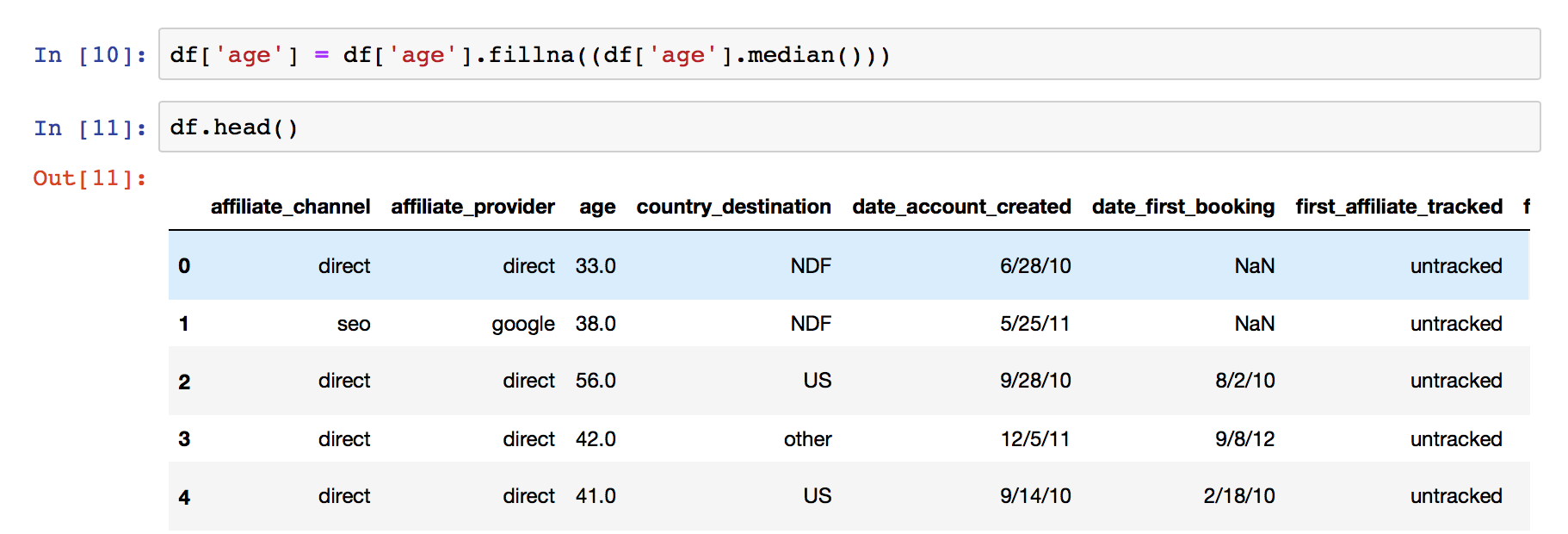


### Date\_account\_created 2

Here we divided the date into three parts, in the form of year, month and day and converted them into numerical. The data after this transformation was used to build all the models again and it actually improved the overall accuracy of each model.



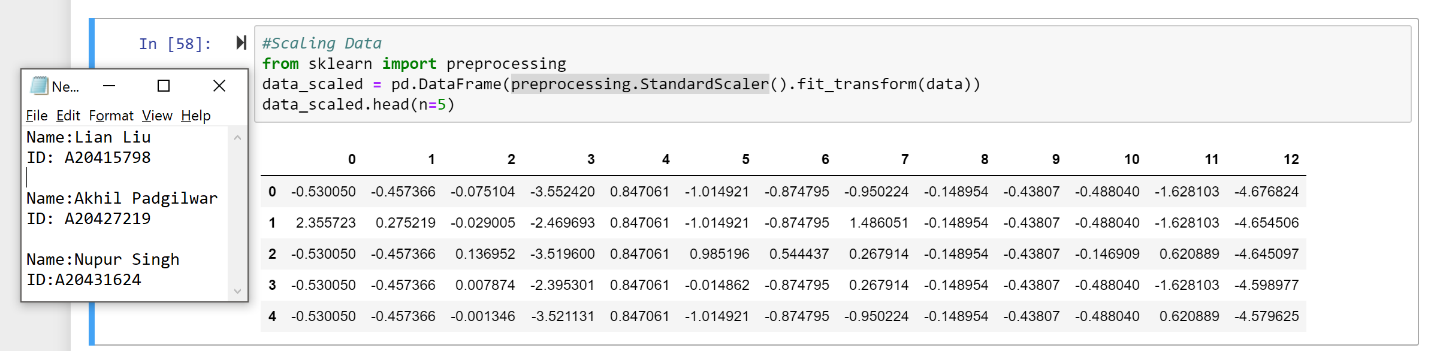
### Replacing of age with median value



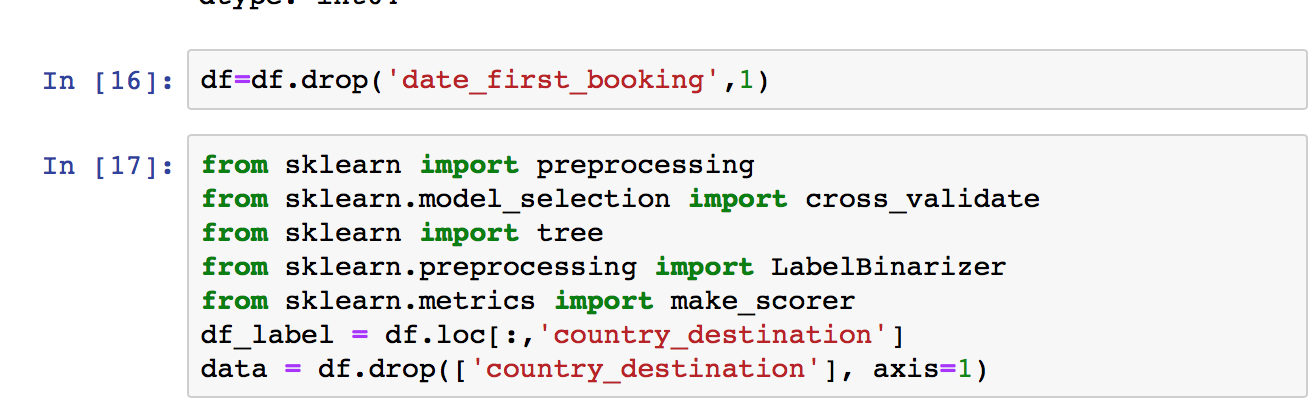
From the exploratory analysis it was quite evident that the age of majority of people who were first time users and booked a country destination were falling under the range of 20-40. Hence we decided to fill the missing values with the median value which was 33.0.

### Normalize(Scaling) Data

After transformation, all the features are numerical feature. However, they stay in a different scale. We need to normalize them into the same scale. We use preprocessing.StandardScaler function to normalize them.



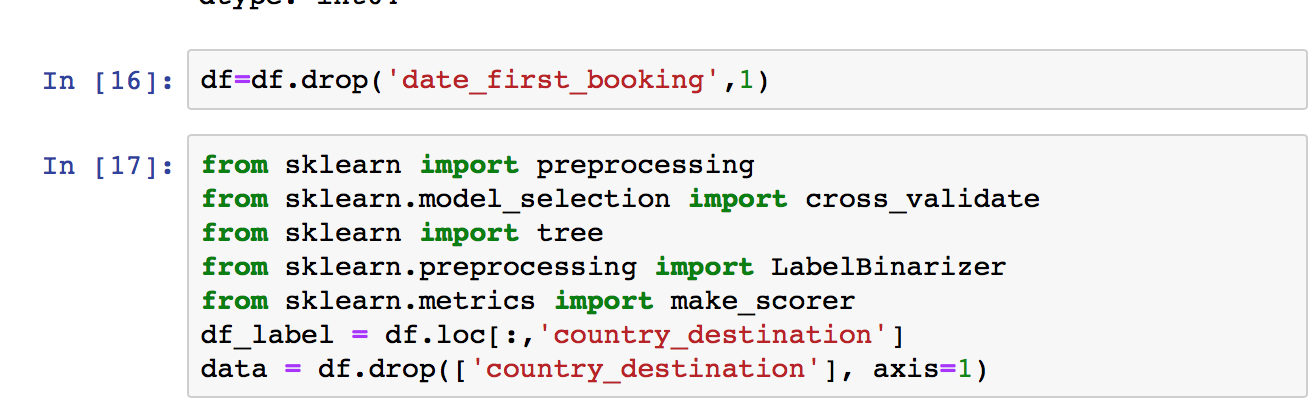
### Dropping the column data\_first\_booking



From the dataset we can see that this column itself had almost 200000 missing values in it which was too much, hence replacing the missing values with something else would have changed the meaning altogether. Hence we decided to drop this feature from our data set.

### Dropping the column id

This column had all the unique id’s of the users. No value was repeating hence we decided to remove this column.



### Converting Labels into numeric



As per the requirement of sklearn library in python, we converted the label into numeric form by first converting it into category and then converting them in the form of codes.

## Model Build

### Naïve Bayes

First, we built a model on preprocessed data but in this data the date field is not preprocessed. We used cross-validation to evaluate the model. We used this data to build the model and accuracy as metrics.

A screenshot of a cell phone

Description automatically generated

From the picture, we are able to know that accuracy is not high. The average accuracy is **67.4%.** Then, we used new preprocessed data having converted date field to build the model.

A screenshot of a cell phone

Description automatically generated

From the picture, we can see that the accuracy is decreased. The average accuracy is **74.3%.**

### K-Nearest Neighbor

The KNN model needs all columns to be numeric and normalized. We have normalized the data to build the model. These two models were built using different evaluation metrics and on before and after preprocessing the date field.

* First model is based on the data which is without the date field preprocessing.

A screenshot of a cell phone

Description automatically generated

From the picture, we can see that the accuracy is **52.08%.**

* Second model is based on the data which is with the date field preprocessing.

A screenshot of a cell phone

Description automatically generated

We can see that the mean accuracy is **74.8%.**

### Decision Tree

For decision tree model, we built two models. We have normalized the data to build a model. These two models were built using different evaluation metrics and on before and after preprocessing the date field.

* First model is based on the data which is without the date field preprocessing. We set max depth is 20. Using Gini and entropy as our criterion.

A screenshot of a social media post

Description automatically generated

From the picture, we can see that the best accuracy is **68.4%**.

* Second model is based on the data which is with the date field preprocessing.

A screenshot of a cell phone

Description automatically generated

From the screenshot, we can see that the best score is **76.8%.**

### Random Forest

In this part, we built random forest models by changing different parameter. For the first model, we set n\_estimators equal to 500 which means we would like to build 500 trees to learn the data. These two models were built using different evaluation metrics and on before and after preprocessing the date field.

* First model is based on the data which is without the date field preprocessing.

A screenshot of a cell phone

Description automatically generated

As we can see, the average accuracy is **70.2%.**

* Second model, we change the n\_estimators as 600. This model is based on the data which is with the date field preprocessing.

A screenshot of a cell phone

Description automatically generated

From the screenshot, we can see that the average score is **79.9%**.

### Support Vector Machines

In this part, we build two SVM model. We use LinearSVC function to build a model. Similar to the SVC with parameter kernel=’ linear’ we implemented the model in terms of liblinear rather than libsvm, so it has more flexibility in the choice of penalties and loss functions and should scale better to large numbers of samples.

* First model is based on the data which is without the date field preprocessing.

A screenshot of a cell phone

Description automatically generated

We can see that the average accuracy is **68.0%.**

* Second model is based on the data which is with the date field preprocessing.

A screenshot of a cell phone

Description automatically generated

The average accuracy is **79.1%.**

### XGBoost

**We build two XGBoost models by using different preprocessed data. XGBoost** is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable.

* First model is based on the data which is without the date field preprocessing.

A screenshot of a cell phone

Description automatically generated

The average accuracy is **70.9%**.

* Next, this model is based on the data which is with the date field preprocessing.

A screenshot of a cell phone

Description automatically generated

From the picture, we can see that the accuracy is **81.43%**.

### Accuracy comparison

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Naïve Bayes | 74.34% |
| K-Nearest Neighbor | 74.87% |
| Decision Tree | 76.83% |
| Random Forest | 79.94% |
| Support Vector Machine | 79.12% |
| XGBoost | 81.43% |

After comparing different accuracies of all the models, we find that **XGBoost** has the highest accuracy which is **81.43%**.

# Conclusion

* Among all the models, XGBoost algorithm works best for our data set.
* Our model can help Airbnb to know exactly where the new user will book their first travel destination.
* It helps users decrease the average time to the first booking.
* Help Airbnb recommend content to new users more effectively and accurately.

# Future Work

* We are able to use Neural Networks to build a model.
* Combination of models
* Using feature engineering like feature selection and feature reduction to select feature.
* Tuning the parameters.