Project Summary

|  |  |
| --- | --- |
| Batch details | PGPDSE – FT Chennai Jul21 |
| Team members | Mohanraj Kandhasamy, Rajendran Karthick Sharan, Yogeshwar Mohan, Dineshkumar Anbalagan, Pydipati Praneeth Kumar |
| Domain of Project | Retail |
| Proposed project title | Price prediction of Airbnb Accommodation |
| Group Number | Group 4 |
| Team Leader | Mohanraj Kandhasamy |
| Mentor Name | Mupidi Srikar |

**Dataset name**

**Introduction to the problem/domain/background details:** The project will be based on the retail domain. In this project we would like to analyze the factors that affect the pricing of Airbnb accommodations and to help hosts to give the best price to the customers and also make good profit in this post pandemic resurgence. We would like to do this with the help of a Machine learning algorithm to find hidden information and relations.

**Problem Statement:** Predicting the price of stay in Airbnb, with the help of a suitable Machine learning algorithm.

**Business problem/ Impact in business of your problem/Need for this study/Abstract (Executive summary):** Since 2008, Airbnb has helped guests and hosts to travel in a more unique, personalized way. The company went from a single air mattress for rent to global cooperation valued at more than 30 billion dollars all thanks to its energetic founder- Brian Chesky. 2020 was supposed to be the golden year for Airbnb as it would go public and issue the worlds’ most sought after stocks. Tragically, Coronavirus happened. The travel sector was gutted by the pandemic. Airbnb now faces burning cash, angry hosts and an uncertain future as 2000 employees could potentially be discharged from their positions and the billion dollars debts with a high-interest rate that is being built to refund their customers. With the help of this project, we would like to help predict the prices of the Airbnb Accommodation based on several factors and help the hosts to quote a suitable price that will satisfy the travelers and to ensure that the hosts with healthy profits, which in-turn will help Airbnb get back on its feet.

**Variable identification**: Independent variables and Target

**Variable information/Data description:**

|  |  |  |
| --- | --- | --- |
| id | integer | Airbnb's unique identifier for the listing |
| scrape\_id | bigint | Inside Airbnb "Scrape" this was part of |
| last\_scraped | datetime | UTC. The date and time this listing was "scraped". |
| name | text | Name of the listing |
| description | text | Detailed description of the listing |
| neighborhood\_overview | text | Host's description of the neighbourhood |
| picture\_url | text | URL to the Airbnb hosted regular sized image for the listing |
| host\_id | integer | Airbnb's unique identifier for the host/user |
| host\_url | text | The Airbnb page for the host |
| host\_name | text | Name of the host. Usually just the first name(s). |
| host\_since | date | The date the host/user was created. For hosts that are Airbnb guests this could be the date they registered as a guest. |
| host\_location | text | The host's self-reported location |
| host\_about | text | Description about the host |
| host\_response\_time |  |  |
| host\_response\_rate |  |  |
| host\_acceptance\_rate |  | That rate at which a host accepts booking requests. |
| host\_is\_superhost | boolean [t=true; f=false] | |
| host\_thumbnail\_url | text |  |
| host\_picture\_url | text |  |
| host\_neighbourhood | text |  |
| host\_listings\_count | text | The number of listings the host has (per Airbnb calculations) |
| host\_total\_listings\_count | text | The number of listings the host has (per Airbnb calculations) |
| host\_verifications |  |  |
| host\_has\_profile\_pic | boolean [t=true; f=false] | |
| host\_identity\_verified | boolean [t=true; f=false] | |
| neighbourhood | text |  |
| neighbourhood\_cleansed | text | The neighbourhood as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles. |
| neighbourhood\_group\_cleansed | text | The neighbourhood group as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles. |
| latitude | numeric | Uses the World Geodetic System (WGS84) projection for latitude and longitude. |
| longitude | numeric | Uses the World Geodetic System (WGS84) projection for latitude and longitude. |
| property\_type | text | Self selected property type. Hotels and Bed and Breakfasts are described as such by their hosts in this field |
| room\_type | text | [Entire home/apt|Private room|Shared room|Hotel] |
| accommodates | integer | The maximum capacity of the listing |
| bathrooms | numeric | The number of bathrooms in the listing |
| bathrooms\_text | string | The number of bathrooms in the listing. |
| bedrooms | integer | The number of bedrooms |
| beds | integer | The number of bed(s) |
| amenities | json |  |
| price | currency | daily price in local currency |
| minimum\_nights | integer | minimum number of night stay for the listing (calendar rules may be different) |
| maximum\_nights | integer | maximum number of night stay for the listing (calendar rules may be different) |
| minimum\_minimum\_nights | integer | the smallest minimum\_night value from the calender (looking 365 nights in the future) |
| maximum\_minimum\_nights | integer | the largest minimum\_night value from the calender (looking 365 nights in the future) |
| minimum\_maximum\_nights | integer | the smallest maximum\_night value from the calender (looking 365 nights in the future) |
| maximum\_maximum\_nights | integer | the largest maximum\_night value from the calender (looking 365 nights in the future) |
| minimum\_nights\_avg\_ntm | numeric | the average minimum\_night value from the calender (looking 365 nights in the future) |
| maximum\_nights\_avg\_ntm | numeric | the average maximum\_night value from the calender (looking 365 nights in the future) |
| calendar\_updated | date |  |
| has\_availability | boolean | [t=true; f=false] |
| availability\_30 | integer | avaliability\_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host. |
| availability\_60 | integer | avaliability\_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host. |
| availability\_90 | integer | avaliability\_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host. |
| availability\_365 | integer | avaliability\_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host. |
| calendar\_last\_scraped | date |  |
| number\_of\_reviews | integer | The number of reviews the listing has |
| number\_of\_reviews\_ltm | integer | The number of reviews the listing has (in the last 12 months) |
| number\_of\_reviews\_l30d | integer | The number of reviews the listing has (in the last 30 days) |
| first\_review | date | The date of the first/oldest review |
| last\_review | date | The date of the last/newest review |
| review\_scores\_rating |  |  |
| review\_scores\_accuracy |  |  |
| review\_scores\_cleanliness |  |  |
| review\_scores\_checkin |  |  |
| review\_scores\_communication | |  |
| review\_scores\_location |  |  |
| review\_scores\_value |  |  |
| license | text | The licence/permit/registration number |
| instant\_bookable | boolean | [t=true; f=false]. Whether the guest can automatically book the listing without the host requiring to accept their booking request. An indicator of a commercial listing. |
| calculated\_host\_listings\_count | integer | The number of listings the host has in the current scrape, in the city/region geography. |
| calculated\_host\_listings\_count\_entire\_homes | integer | The number of Entire home/apt listings the host has in the current scrape, in the city/region geography |
| calculated\_host\_listings\_count\_private\_rooms | integer | The number of Private room listings the host has in the current scrape, in the city/region geography |
| calculated\_host\_listings\_count\_shared\_rooms | integer | The number of Shared room listings the host has in the current scrape, in the city/region geography |
| reviews\_per\_month | numeric | The number of reviews the listing has over the lifetime of the listing |

**Future Work/Methodology (Details of algorithms):**  Once this process has been implemented we can collect the reviews of the customers and perform sentiment analysis using NLP. We can also deploy the price prediction algorithm.

**Timeline Chart (Weekly plan):** Tentative weekly plan that you will be following.

|  |  |  |
| --- | --- | --- |
| WEEK 1 | 21-11-2021 | SYNOPSIS - Understanding the Dataset and Business Aspects, Performing |
| WEEK 2 | 28-11-2021 | EDA and various Feature engineering techniques on dataset |
| WEEK 3 | 02-12-2021 | Interim presentation & Report |
| WEEK 4 | 12-12-2021 | Working Progress Status I |
| WEEK 5 | 19-12-2021 | Working Progress Status II |
| WEEK 6 | 30-12-2021 | Final Report |
| WEEK 7 | 31-12-2021 | Final Presentation |

**References (Data set source/Journals/articles)**

**Data set Source:** https://public.opendatasoft.com/explore/dataset/airbnb-listings/table/?disjunctive.host\_verifications&disjunctive.amenities&disjunctive.features

**Declaration: This is to declare that the dataset that we are using for our capstone project does not have any relevant legality associated to it** **and can be used to showcase the work we do on it as a presentation in Great Learning.**