CSCE 5310: Methods in Empirical Analysis

Project Proposal

Project Name: Airbnb price predictions

Team Members:

1. Roshan Sah [11574385]
2. Saisri Teja Pepeti [11555656]
3. Yamuna Bollepalli [11552426]

INTRODUCTION

People today struggle to locate the greatest locations for gatherings with family or friends also the present trend of destination wedding spots.

A few people had the idea to create an online platform where people could interact with various regions who offer the best rentals in the most genuine and pleasant locations for casual parties, personal meetings, and get-together parties.

The primary benefit of Airbnb is that it exclusively partners with hosts that offer superior service to their guests.

The majority of individuals will look for locations with the best architecture.

In this aspect, we are so eager to find how Airbnb bookings are affected by the variations in prices and traffic in the area.

Goals and Objectives:

- Main goal is to predict the places where most of the people are booking the places and various reviews of the places

Motivation

- To observe If there are any observable differences in traffic across different areas,what might be the cause of those differences?
- To predict change in prices based on the location and reviews and also predict the busiest airbnb hosting areas.

Significance

- Importance of this project is to give the better price updates abruptly as per the locations and the traffic information for the customers to navigate to other options quicker.
- Also prompt updates about the busiest hosts and the availability

Literature Survey

We have referred to relevant research papers, on how we can implement the validation methods on huge and known fields of data sets also which have more information as data to classify and validate the test scores for more accuracy against the independent variables to dependent variables. By going through some case studies like movies box office predictions, medical charges prediction and amazon fine food review analysis. Upon checking all these case studies we were lacking some information as data inputs. Thus, we came across this interesting case study considering current covid situations how the market is going through the hotel and rental services. Thereafter we thought to bring certain ;outcomes by taking location based analysis for those who are looking for immediate services with the accuracy depending on the demand and the hosts business factors.

Background work references of case studies:

1. *https://www.researchgate.net/publication/337550899_Predicting_Movies%27_Box_Office_Result_-_A_Large_Scale_Study_Across_Hollywood_and_Bollywood*
2. *https://journalofethics.ama-assn.org/article/challenge-understanding-health-care-costs-and-charges/2015-11*
3. *https://snap.stanford.edu/data/web-FineFoods.html*

Features

- After Data cleaning and filtering, we'll retrieve the data and identify the relation between the variables and will also visualize the data.
- Mainly we will use Hypothesis-testing, T-test for validation.
- We're planning to represent various score values like R2 score, AUC score etc.

Dataset:

We will take the dataset which contains almost 45k rows with 16 columns for each. It is composed of 3 float types, 7 int types and 6 object types of id, name, host_id, host_name ,neighbourhood_group, neighborhood, latitude, longitude, room_type, price minimum_nights, number_of_reviews, last_review, reviews_per_month calculated_host_listings_count,availability_365.

Pre- Processing:

For the pre- processing, we will perform the data wrangling and cleaning for merging complex data sets and correcting errors to make them more accessible and understandable.

Model:

We will use the Linear Regression models: Simple linear and multi linear regression models to predict the dependent value prices of Airbnb's with help of all other independent variables associated with it like reviews, locations and bookings and availability…

Also we will use validation methods, hypothesis testing and T-Test to see if Airbnbs have been open for at least more than a month, the sample availability differs significantly from the total average availability of all Airbnbs.

Suppose if we even take a sample of the availability data, the average Airbnb availability would remain the same which can be said for and do for Null Hypothesis

Alternative: If we take a sample of the available data, the average Airbnb availability may come different.

Expected outcome

We are planning to use linear regression for finding the efficiency of our model.

Our expected final outcome will be the accuracy of predicting the airbnb prices. We will be Implementing various models to check which model is performing better and giving more accuracy.

References

Peilu Liu -

https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1979&context=etd_projects