MATH301Project

Reese Guo

April 2020

1 Dataset Introduction

The data of interest is the New York City Airbnb data. The Airbnb dataset describes the Airbnb listing activities and metrics in New York City in 2019. The dataset was open-source and was obtained from Kaggle [1]. The dataset has 48878 records and 16 columns. Each record in the dataset stands for a property listing on Airbnb. Information about each record include listing ID, name of the listing, host ID, name of the host, location of the listing in large neighborhood areas, latitude, longitude, listing property type, price in dollars, amount of nights minimum to book, number of reviews, latest review date, number of reviews per month, amount of listing per host, number of days when listing is available for booking. For this project, we will consider the number of days when a listing is available for booking to be our sensitive information and will try to synthesize this variable based on the neightborhood areas, listing property type, and number of reviews information.

2 Research Question

The research question of this project is how to effectively synthesize the number of days when a listing is available for booking based on other continuous and categorical data of the Airbnb dataset. This project chooses to use Probit Regression and to investigate the effectiness of Probit Regression by checking the utility and evaluating the disclosure risk of the synthesized dataset. Although the Airbnb dataset we are using for this project is already released, we will use it as if it is unreleased data to explore the effectiveness of the synthesis method.

3 Background and Significance

Airbnb, as a leading sharing lodging service provider, quickly gained popularity among tourists in recent years. As Airbnb rapidly expands, security of both property renter and property host becomes a concern. A previous study has found that there is a positive correlation between the spatial distribution of Airbnb and the number of property crimes [2]. This result suggests that an

Airbnb property is more prone to property crimes than non-rental properties. Thus, when releasing data about Airbnb listings, we should protect the information of how often a property is listed for rental to protect a property from potential property crimes. This attribute is represented by the number of days a listing is available for booking is the New York City Airbnb dataset. Therefore, we will focus on the synthesis of this variable in this project.

A success synthesis model can not only protect privacy for listing hosts, but also preserve the data accuracy and the correlation of the synthesized variable with other variables so that other data analyst can perform analysis on the dataset without losing valuable information.

4 Methods

We will first divide the number of days a listing is available for booking into 6 categories. The reason we converted this continuous variable into a categorical variable is that if we have property A with 340 days available for listing and property B with 350 days available for listing, there is no meaningful difference, to our concern, between A and B, since they are all fully rental properties. The converted categorical variable can accurately capture the information the main use of a property, whether being mainly rental or being mainly private use. However, the drawback of this approach is also obvious. For records whose number of available days variable are right around the category cut-off points, the converted categories may not be very representative since there is no significant difference between those records and records in their neighboring category. A mitigation solution to this problem is to create more categories to reduce the differences between neighboring categories, while maintaining the meaningfulness of the categories. Thus, we decide the number of days a listing is available in a year into six categories, with each category containing 60 days and the last category containing 65 days.

After converting the number of days available for listing variable into categories, we will use probit regression to fit the observations. The probit regression can be expressed as follows,

$$\begin{aligned} \epsilon_1, ..., \epsilon_n &\sim i.i.d\ normal(0, 1) \\ Z_i &= \boldsymbol{\beta}^T \boldsymbol{x}_i + \epsilon_i \\ Y_i &= g(Z_i), \end{aligned}$$

where Y_i is the number of available days observation for the i^{th} record, $\boldsymbol{x}_i$ is the vector of predictors for the i^{th} record, and $\boldsymbol{\beta}$ and g are unknown parameters specific for the regression [3]. We will assume that the availability of a property is related to the neighborhood it is in (denote later by Neigh), the room type of the property (denote later by Room), and the number of total review it has

(denote later by Review). Thus, the predictor $\boldsymbol{x}_i$ can be expressed as,

$$\boldsymbol{x}_i = (Neigh_i, Room_i, Review_i, Neigh_i \times Room_i, \\ Neigh_i \times Review_i, Room_i \times Review_i).$$

A Gibbs sampler, with 5000 iterations, was written in R to simulate the parameters for the probit regression. Then, synthesized data were generated using simulated parameters. Upon obtaining the synthesized dataset, the utility of the dataset was measured by calculating the propensity score [4] and the disclosure risk of the synthesized data was measured by calculating the expected match rate, the true match rate, and the false match rate of the synthesized dataset [5].

References

[1] Dgomonov. New York City Airbnb Open Data, 2019. `https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data`, Accessed on 2020-04-20.

[2] Yu-Hua Xu, Jin-won KIM, and Lori Pennington-Gray. Explore the spatial relationship between airbnb rental and crime. 2017.

[3] Peter D Hoff. *A first course in Bayesian statistical methods*, volume 580. Springer, 2009.

[4] Mi-Ja Woo, Jerome P Reiter, Anna Oganian, and Alan F Karr. Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, 1(1), 2009.

[5] Jingchen Hu. Bayesian estimation of attribute and identification disclosure risks in synthetic data. *arXiv preprint arXiv:1804.02784*, 2018.