MATH301Project

Reese Guo

April 2020

1 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. We choose to use probit regression and to investigate the effectiveness of probit regression by checking the utility and evaluating the disclosure risk of the synthesized dataset. A synthesized dataset's utility refers how well the synthesized dataset preserve the distribution and relationships among variables in the original dataset. High utility suggests that the results of statistical analyses performed on the synthesized dataset are close to the results if the analyses were done on the original dataset. Disclosure risk refers to the risk that data intruder, with partial information of the dataset, correctly infer sensitive information about a record in the dataset. A low disclosure risk is usually required for synthesized datasets for privacy protection. 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.

2 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 [1]. 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 utility 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.

3 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 [2]. 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.

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. Thus, we decide to divide 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 model the observations. Probit regression is commonly used to model ordered categorical variables outcomes with specified predictors. The probit regression can be expressed as follows,

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

where Y_i is the category of number of available days 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. Our description of the probit regression follows closely to [3]. We will assume that the availability of a property is related to the neighborhood it is in, the room type of the property, and the number of total review it has. The variable names used by this project and their meanings are presented in Table 1.

Variable Name	Variable Meaning
$Neigh_i$	the neighborhood the i^{th} record property is in
$Room_i$	the room type of the i^{th} record property
$Review_i$	the log transformed total amount of review of the i^{th} record property.

Table 1: Variable Description

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 and the disclosure risk of the synthesized data was measured by calculating the expected match risk, the true match rate, and the false match rate of the synthesized dataset [4]. Propensity score is a common method used to estimate the effect of a treatment, policy, or intervention between two groups [5]. In our case, we measure the effectiveness of the synthesis process. Expected match risk, true match rate, and false match rate are three summarizing aspects of identification disclosure risk for a dataset. Expected match risk measures how likely it is to find the correct match for each record and for the whole dataset. True match rate measures the percentage of true unique matches. False match rate measures the percentage of unique matches to be false matches. [5]

When evaulating the identification disclosure risk, we assume that the intruder has knowledge of the neighrborhood group, the room type, and the listed price for a property since these information are available to the public on Airbnb websites when booking a property.

5 Results

A Gibbs sampler with 5000 iterations was run to generate the posterior draw of β, g and Z of the probit regression. The 5000^{th} iteration β, g and Z values were taken to synthesize the new dataset. Synthesized categories of number of available days are then plotted with the original categories in Figure 1. As shown in Figure 1, the synthesized data generally follows the same distribution as the original dataset, with some differences in category $4, 5,$ and 6.

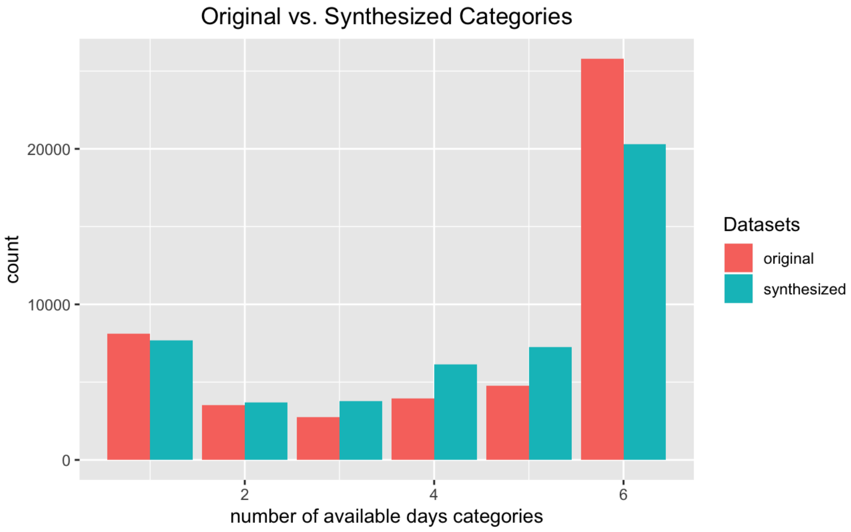

Figure 1: Barplot of original and synthesized categories

To quantitively measure how well the synthesis dataset preserve the properties of the original dataset. we calculated propensity score to measure synthesized dataset's utility. A propensity score closer to 0 suggests high similarity between the synthesized and original datasets and a propensity score closer to $\frac{1}{4}$ suggests low similarity between the synthesized and original datasets. The propensity score calculated for the synthesized Airbnb dataset is,

$$U_p = 0.0005568.$$

The calculated propensity is close to 0, which suggests high similarity between the synthesized and original Airbnb datasets and thus a high utility for the synthesis process.

The results for the identification disclosure risk measurement are as follows,

$$\begin{aligned} s &= 30 \\ \text{EMR} &= 80.39 \\ \text{TMR} &= 0.00020 \\ \text{FMR} &= 0.67, \end{aligned}$$

where s is the number of unique matches in the synthesized data, EMR is the expected match risk, TMR is the true match rate, FMR is the false match rate. Since the dataset has 48878 records, the expected match rate of 80.39

indicates that the probability for each record in the synthesized dataset to be correctly identified is $\frac{80.39}{48878} = 0.0016$. The true match rate of 0.00020 indicates that $0.00020 \times 48878 \approx 10$ records are correct unique matches. Lastly, the false match rate of 0.67 indicates that among the 30 unique matches, $30 \times 0.67 = 20$ are false matches. Overall, the identification disclosure risk for the synthesized Airbnb data seems very low, indicating a high level of confidentiality protection.

6 Dicussion (place holder, please ignore its content in this draft)

We categorized the continuous variable avialble days into categories when analyzing the data. However, this approach has drawback. 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.

References

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

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

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

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

[5] 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.