MODS207 Applied project

Estimating gender and race of AirBnB guests

Students: Thang Bui Doan

Palaiseau, July 2023

1. Introduction

The global outbreak of COVID-19 has had a profound impact on economies worldwide, and the tourism and hospitality industry, including companies like Airbnb, has experienced significant disruptions. The pandemic brought travel restrictions, lockdown measures, and a decline in consumer confidence, leading to a sharp decline in tourism and travel demand. Apart from that, the COVID-19 also caused a wave of discrimination towards Asian-Americans [1]. In order to adapt to the circumstance, AirBnB adapted numerous policies, such as flexible cancellations, promoting local tourism, detecting and preventing discriminations. To accomplish the last goal, the first step is to be able to predict a person's gender and ethnicity.

In this project, we use available methods to predict gender and ethnicity of AirBnB guests from available provided data. I hope the result of this project can help us resolve part of the problem, which is gender and race discrimination, whether the world is plague with a epidemic or not.

2. Background information

Airbnb is a globally renowned online marketplace that revolutionized the way people travel and find accommodations. Founded in 2008, Airbnb provides a platform where individuals can list, discover, and book unique accommodations in over 220 countries and regions around the world. With a diverse range of options including apartments, houses, villas, and even unconventional spaces like treehouses and yurts, Airbnb offers travelers an alternative to traditional hotels, providing a more personal and immersive experience. By connecting hosts and guests directly, Airbnb promotes a sense of community and cultural exchange while empowering individuals to monetize their extra space and discover new opportunities. Today, Airbnb is not just a platform but a global travel community that has forever changed the way people explore and experience new destinations.

Solving the problems of discrimination may help AirBnB provide their customers with better services, become a bigger, leading companies in the tourism industry, starting with the prediction of their guest genders and ethnicities.

3. Data collection strategy/Algorithms

3.1. Data collection.

The provided dataset was scrapped through AirBnB website, contains records of guest stays at AirBnB houses. Each sample is record of guest information, also the host review about them. The which contains 10000 samples, corresponding to 10000 stays.

Each row of the dataset has the following field:

- Listing_id
- Guest_id
- Review_date:
- Name: Name of the guests
- Profile
- Address:
- Language: Language the guests
- Job: Profession of the guests
- Review: Review of the house owner to the guest
- Image: Link to download picture of the guests.

Since our objective is to predict a person's gender and ethnicity we would only keep the name and the image of the guest, which would be downloaded from the internet.

To make the prediction process simpler and the prediction process be more precise, based on the name field of the data, I excluded samples that have more than 2 guests (such as: Hank & Julie, Brittany + Brandon, …) and samples which pictures are missing.

Moreover, since there is no information about the actual gender and ethnicity, the labels of these values would need to be filled out manually.

Based on the name and downloaded images, the actual gender of a guest can either be m (male) or f (female), the ethnicity value can be either w (white) or o (other), or na (Not available, this value is filled for picture with low resolution or there are more than one person in the picture, or picture does not contain any visible human, the angle of the face or lightning does not support the identification).

3.2. Algorithms.

In this project, to predict gender and ethnicity of a person, we use these following tools: Genderize.io, NamePrism and DeepFace.

Genderize.io is an API that utilizes machine learning algorithms to predict the gender associated with a given name, enabling developers to incorporate gender analysis into their applications.

NamePrism, on the other hand, is a tool that uses statistical modeling to estimate the gender distribution associated with names across different countries and cultures, offering insights into naming patterns worldwide.

Through the usage of NamePrism API, we collect the data in JSON format. NamePrism divide people ethnicity into 6 categories: White, Black, API (Asian and Pacific Islander), AIAN (American Indian and Alaska Native), 2PRACE (more than 2 race) and Hispanic. However in the limitation of this project, we will rearrange these races into 2 groups: White and other.

Lastly, DeepFace is a sophisticated facial recognition system developed by Facebook's AI research team. It supports various applications: face verification, face recognition, face detectors and face analysis.

In order to achieve our goal, prediction of gender and ethnicity, we rely on 2 functions of DeepFace: face recognition and face analysis. DeepFace supports the first function through different backends: opencv, retinaface, dlib, … After recognizing face of a person, the isolated region is passed directly to the analysis functions, which will return with prediction of gender and ethnicity. From the 6 probable output races of DeepFace (which are: Asian, Indian, Black, White, Middle Easter, and Latino Hispanic), we are gonna regroup these races into White and other.

4. Data description and analysis

4.1. Data description.

Because of time-consuming labeling process, the remaining dataset has 1028 samples.

In which, there are 600 people labeled as female, 428 labeled as male, 26 pictures that are not a selfie of the guest or cannot be recognized easily. These 26 pictures account for 2.5% of overall 1028 pictures, which is considered a small amount.

Gender	Count	Percentage
Male	428	41.63%
Female	600	58.37%
Total	1028	100%

Regarding ethnicity of a guest, here are the statistic information:

Ethnicity	Count	Percentage
White	687	66.83%
Other	287	27.92%
NA	54	5.25%
Total	1028	

Overall, white ethnicity makes up the majority, about two thirds of 1028 samples. The NA value only has 54 samples, which represents about 5% of the dataset.

4.2. Analysis.

4.2.1. Prediction of gender.

In case we only consider name of the person, Genderize.io achieved an accuracy of 93%, which was high.

Here is the confusion matrix of the prediction, also precision and recall for each label:

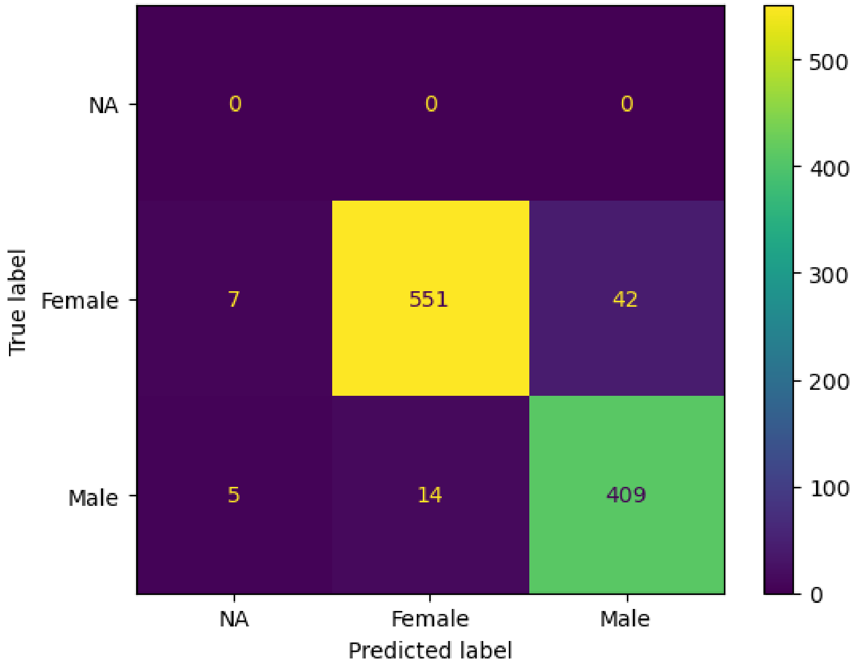

Figure 1: Confusion matrix of gender prediction with Genderize.io

Label	Precision	Recall
Male	90.69%	95.56%
Female	97.52%	91.83%

Table 1: Prediction result with Genderize.io

All the precision and recall score of female and male label are above 90%. However, the female gender has higher precision. On the other hand, the male gender has better recall. There are also 12 names that the algorithms can't predict, they are all pronunciation of foreign name in Latin characters.

We utilize DeepFace algorithm to predict gender of guests from their pictures. In this project, I have tried three backends: opencv, retinaface and dlib.

Backend	Not detectable	Accuracy
OpenCV	55.54%	17.02%
RetinaFace	5.16%	70.72%
Dlib	19.16%	63.91%

Table 2: Gender prediction result with DeepFace

Out of the three backends, RetinaFace performed best, next is Dlib and OpenCV performed worst.

As we can see on the table above, OpenCV could not detect human face in 55.54% of all cases - 571 out of 1028 pictures. Even when we consider there are 26 pictures with more than 1 person, or there is no person at all, OpenCV backend performed very poorly.

RetinaFace has best results, its accuracy rate is 70.72%. However, the backend misclassify a lot of women into male gender.

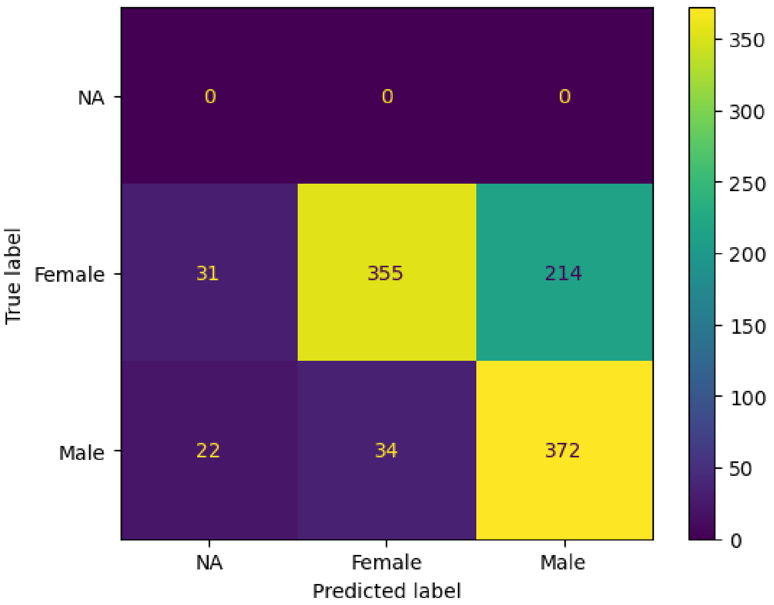

Figure 2: Confusion matrix of gender prediction with RetinaFace backend

Using pictures seems to be much less efficient compared to the solution of using name only. This can be explained that there are a lot of pictures that have more than one person. Although through the guests name, samples that may included more than one person had been filtered out, reading the review column of the original dataset suggests there much more samples that contain more than one person but the name field only has name of one.

4.2.2. Prediction of ethnicity.

After excluding all samples that we could not label their ethnicity, the dataset has 974 samples left.

We have 2 ways to predict a person's ethnicity: through their names and pictures.

The first direction is to use NamePrism. The solution has accuracy of 71.35% and macro F1 of 56.98%. As we can see in the following confusion matrix, the white label has a much higher recall rate, compared to other races. On the other hand, the result on "other" label is much worse. This could be explained that a lot of people that have foreign origin may not use name from their countries but adopted a local name instead. So using name alone is not a good way to predict their ethnicities.

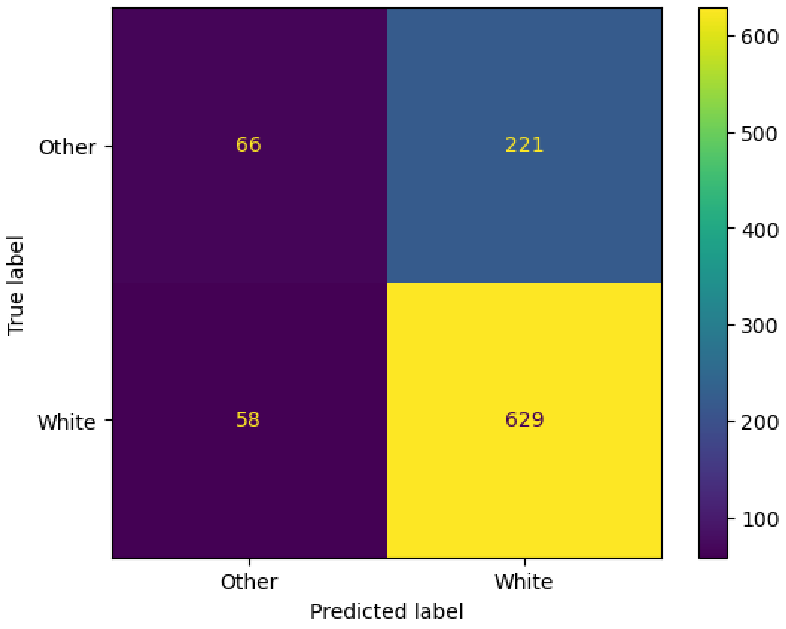

Figure 3: Confusion matrix of ethnicity prediction using NamePrism

Moving on to ethnicity prediction using DeepFace, we have the results of 3 backends:

Backend	Not detectable	Accuracy
OpenCV	26.59%	58.32%
RetinaFace	1.54%	80.29%
Dlib	15.09%	71.77%

Table 3: Ethnicity prediction result using DeepFace

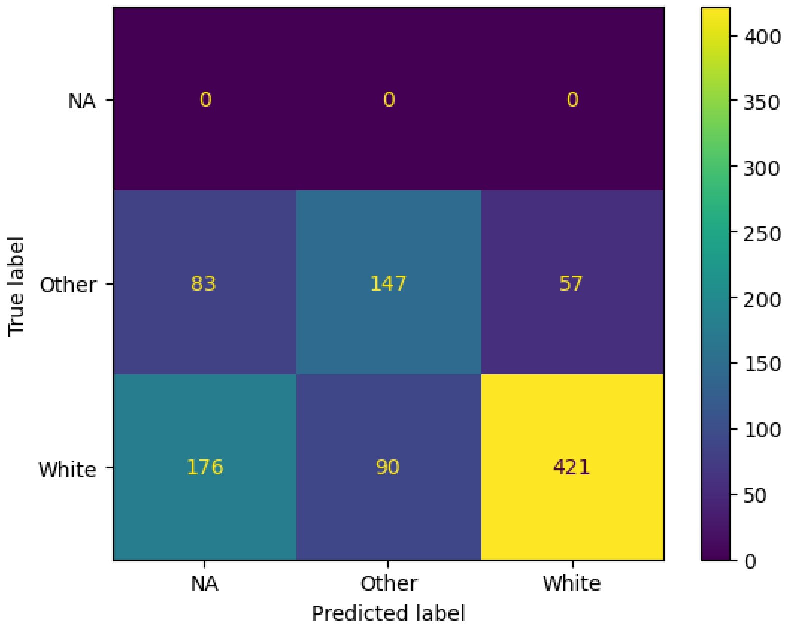

Figure 4: Confusion matrix of ethnicity prediction using OpenCV

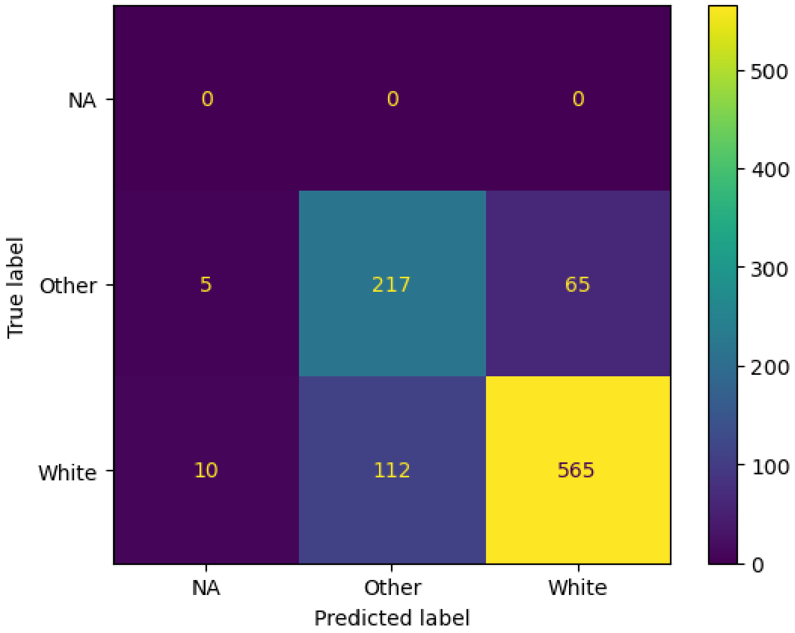

Figure 5: Confusion matrix of ethnicity prediction using Retina Face backend

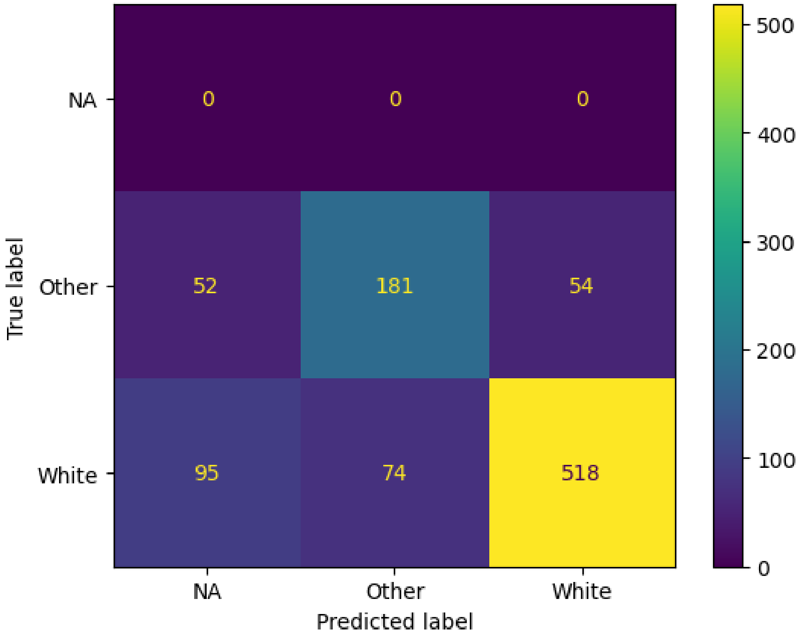

Figure 6: Confusion matrix of ethnicity prediction using Dlib backend

The RetinaFace backend remains to be the best, since it managed to detect the most number of face in all three backends. It achieved the accuracy of 80.29%

On the other hand, the OpenCV has the worst performance, its accuracy is only 58.32%. Dlib has accuracy of 71.77%, which is lower and equal to the accuracy of NamePrism, respectively.

5. Conclusion

Overall, the gender prediction process have much higher accuracy compared to ethnicity prediction. Using only name in gender prediction already brings back high result. The solution to use pictures proved to be less efficient and more time-consuming. A reason for this can be that sometimes there are more than 1 person in the picture, the picture does not have a good resolution, angle or lightning. We can improve the accuracy of DeepFace by filter out the pictures that have more than 1 person, or program the algorithm to choose the person in the center of the picture, incase of detecting multiple faces.

On the other hand, predicting ethnicity seems to be much harder topic. The best solution is to use DeepFace, with Retina Face backend. And since we resource to using pictures, this solution have the same problems that we have stated above. Moreover, both DeepFace and NamePrism also supports prediction of multiple classes. Retrain these two models to predict two labels only (White and Other) might bring good results.

Finally, this project has not resolve the problem of using both people names and pictures to predict their genders and ethnicities, only using one of them at a time. So in the future, I would like to continue this project to try combining the two methods, finding a better a solution that what we already achieved.