An Assessment of the Effect of Race and Gender on Police Service Satisfaction in 2012

Abstract

This paper intends on assessing the reliability of using Logistic Regression to build a model that would predict the satisfaction of police service quality in Arlington County, Virginia in 2012. The model makes use of two features, gender, and race, of the residents who took part in the survey. The model then aims to identify the level/class of satisfaction of participants based on these two features. The fitness of the model is then considered using various measures of accuracy.

1. Introduction

The Black Lives Matter movement was a recent demonstration that highlighted police brutality and brought to light that one's opinion on police services was linked to their race, which would theoretically make it easy to discern the police's treatment of that race as a whole today.

Based on the research outlined in a 2020 paper titled "What matters in citizen satisfaction with police: A meta-analysis," most studies show that Blacks hold a very different perspective on the police system than Whites. That is, police are found to be "less favorable" in their eyes.

However, it is worth noting that some studies do not indicate that any relationship between race and the opinion of police exists. Nonetheless, with the recent Black Lives Matter movement and initiatives to defund that police, it can confidently be said that the view of police has decreased exponentially in today's world, especially by those who feel the effects of the police system the most.

In regards to gender, predicting police service satisfaction is a bit more nuanced. Some studies show that women have a more favorable opinion of the police, while others show the correlation between gender and "Police Service Satisfaction" is insignificant.

However, since gender is such a defining factor in one's choice in many other facets of one's life, we felt this was a feature worth looking at.

Our analysis and models specifically work to see if the former point holds true for people nearly a decade ago.

2. Data Cleaning and Preliminaries

Our data was retrieved from data.gov and consisted of a Microsoft Excel spreadsheet that cataloged questions given to residents of Arlington County, Virginia in 2012 regarding various aspects of the city and its services. Each question was designated one column, although questions that were composed of several subquestions were spread out over multiple columns.

The survey was responded to by 1,306 participants and appears to have thirty-five questions; this does not include the components of questions that have more than one part.

As our analyses really only require information regarding each participant's race and gender, after we converted the Excel spreadsheet into a Python

pandas-ready CSV file, we removed all the features that did not include any data about these two. In the end, we were left with two pieces of information regarding each participant: their race and gender.

While these were the only two features that remained, we noticed that there were more than six different race options given in the survey and each of them were split into different columns. As we wanted to look at the race feature as a whole, we realized we had to have one homogenous column depicting the participants' races. To do this, we merged all the race-related columns.

However, upon doing this, the participants who had identified as more than one race had *their* races essentially merged, as well. For example, if a participant had identified as "White" and "American Indian/Eskimo," they would now be referred to as the race "WhiteAmerican Indian/Eskimo," for we were not at liberty to ask the participant which of the two races they would consider themselves to identify more with so we could place them in simply one category.

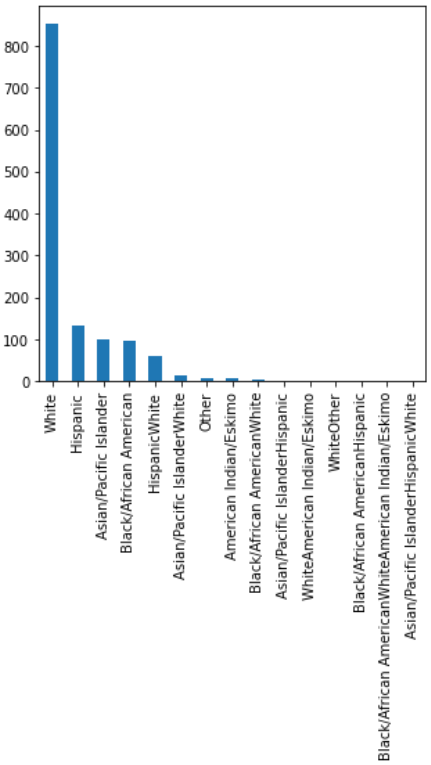

Figure 1 A histogram depicting the frequencies of the different races present in the data set *before* cleaning

It is clear that many of the races that are on the right side of the histogram have *very* low, near inconsequential frequencies. Because these data points did not seem significant enough for us to include in our analysis, we decided to remove the data of any participants who identified as a combination of races or who identified as "Other." After this removal process, we were left with a final set of race data that consisted of a total of five different races.

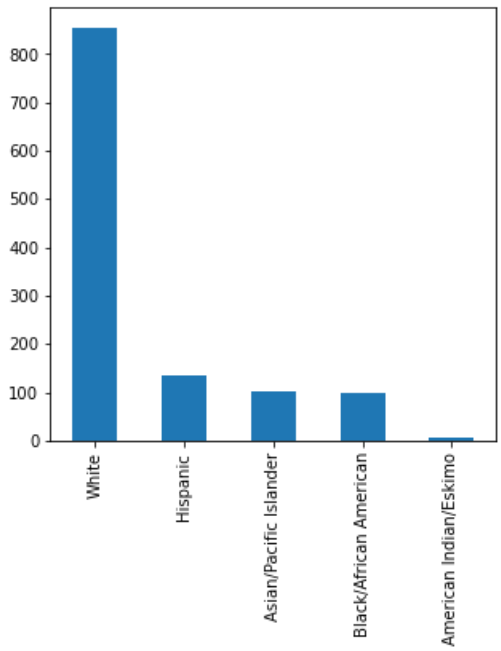

Figure 2 A histogram depicting the frequencies of the different races present in the data set *after* cleaning

Because we were also including gender in our analysis of race and police service satisfaction, we had to ensure that there was a significant number of data points in our gender column, as well.

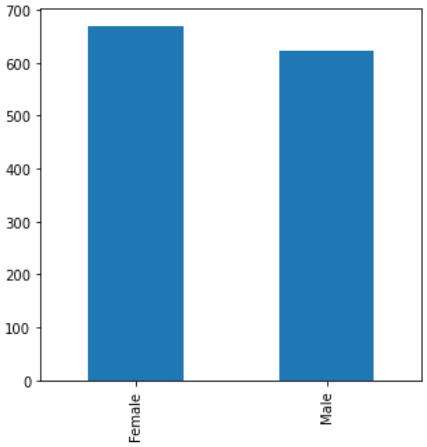

Figure 3 A histogram depicting the frequencies of the different races present in the data set

We were satisfied with the distribution between male and female participants, so we saw no need to further manipulate or restructure the data in regards to the gender feature.

We decided at that point to move onto our class—that is, police service satisfaction.

After employing the Python pandas `value_counts()` method, we discovered that there were six different levels of satisfaction of police services in the survey: Very Satisfied, Satisfied, Neutral, Dissatisfied, Very Dissatisfied, and Don't Know.

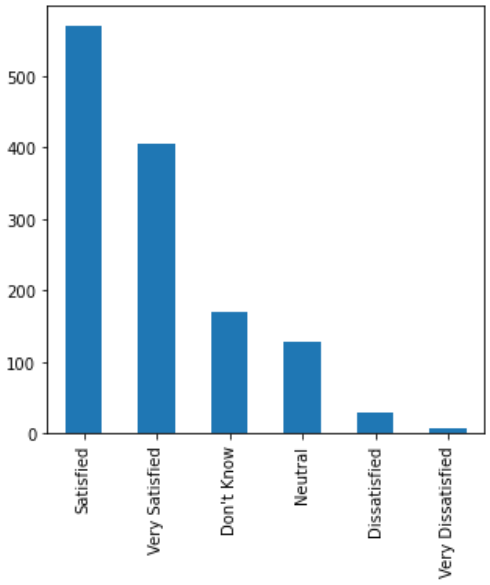

Figure 4 A histogram depicting the frequencies of the different levels of satisfaction present in the data set before cleaning and making the class binary

In order to make our class a binary class, we decided that it would be most beneficial to create two extremes of the satisfaction class: Satisfied and Dissatisfied. However, to expand our class data size, we considered the fact that the difference between a person that is "very satisfied" with the police service should not see too much of a difference when compared to a person that is simply "satisfied" the police service; as such, we decided to combine those two levels to create the Satisfied half of our binary class.

The same thing was done to those who were “very dissatisfied” and “dissatisfied” to create the Dissatisfied half. Since the “Don’t Know” level is essentially useless from an analysis standpoint, we removed the data points of participants who answered as such. And since “Neutral” is not one of the two extreme levels of satisfaction that we would like to have our model look at, we also removed those data points, as well.

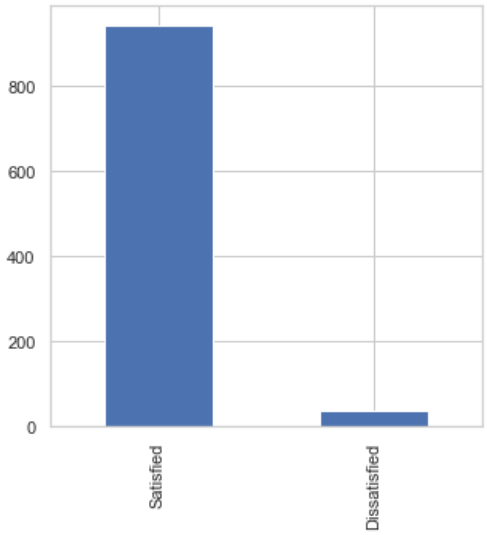

Figure 5 A histogram depicting the frequencies of the different levels of satisfaction present in the data set after cleaning and making the class binary but before balancing

By looking at this histogram, it’s plain to see that our "Police Service Satisfaction" class is *extremely* imbalanced, and it’s important to have balanced data in order to have usable results. Since our Dissatisfied satisfaction level is significantly lower than our Satisfied level, we decided to upsample our Dissatisified satisfaction level (after having split it into training and testing data) so that then we would have a more balanced collection of data points.

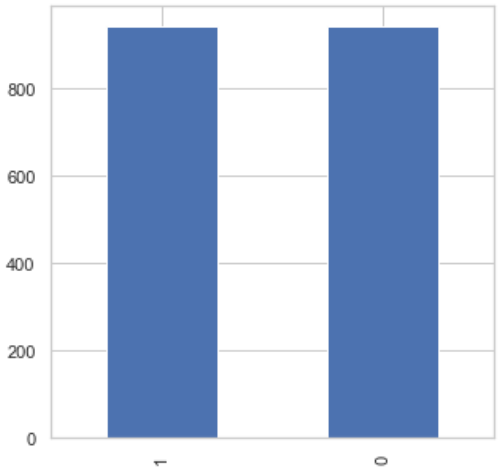

Figure 6 A histogram depicting the frequencies of the different levels of satisfaction present in the data set after cleaning and making the class binary and after balancing; 0 represents “Satisfied,” whereas “Dissatisfied” is represented with 1

3. Experiment

Finally, for ease of data processing, we assigned numbers to each item in our table. It has been alluded to in our previous histogram, but in our police service satisfaction class, everywhere a participant had said “Satisfied,” we changed their English-text response to the number “1” and everywhere they had written “Dissatisfied,” we changed their response to the number “2.” Similarly, in the gender observation column, “Female” was changed to the number “0” and “Male” was changed to the number “1.” As we ended up looking at five different races in the race observation column, we assigned each different race a number ranging from “0” to “4.” The assignments were “White” became “0,” “Hispanic” became “1,” “Asian/Pacific Islander” became “2,” “Black/African American” became “3,” and “American Indian/Eskimo” became “4.”

Following the extensive process of cleaning, we then proceed to assign features and labels to both our variables, *X* and *y*, split the data, and finally fit the data to a logistic regression model. Naturally, we fixed the features “Gender” and “Race” to our *X* variable and "Police Service Satisfaction" to our *y*. We performed a

default 75% and 25% split to divide our cleaned data *before* balancing (mentioned in the earlier section). Performing balancing of any kind after splitting the data ensures that additional observations are not added to the data used to test the model. This prevents the testing data from being skewed and produces the best measures of fitness. We then proceed to fit the model using Logistic Regression.

This classification method was chosen over a regression method as satisfaction levels could clearly be defined into separate classes. More specifically, the satisfaction levels were discrete values rather than continuous. The question of choosing Logistic Regression or Naive Bayes for classification arose. Ultimately, it was decided that Logistic Regression would be best as the dataset contained a large number of samples.

Using sci-kit learn's Logistic Regression modeling, the train data was fit to the model and the model was assessed.

4. Discussion

We first compared the classification outputs from predicted versus actual data. Actual values appeared in the order of 1, 1, 1, 1, 1 while the respective predicted values were 0, 1, 0, 0, 0. Based on this alone, it is clear that the model correctly predicts satisfaction *very* infrequently.

We then looked at the accuracy and F1 score of the model. After multiple runs of the model, the highest accuracy score we could achieve was around sixty-one percent. This score shows that the model accurately predicts satisfaction over half of the time, but still predicts inaccurately nearly forty percent of the time. While this score is not optimal, it has some degree of reliability. Similarly, our F1 score was roughly sixteen percent, which is quite unfortunate. When a confusion matrix is generated, the results mirror our accuracy score and indicate a poor model.

Finally, the provided heatmap shows the correlation between race, gender, and police service satisfaction. The numbers are incredibly low which is expected based on our accuracy and F1 score.

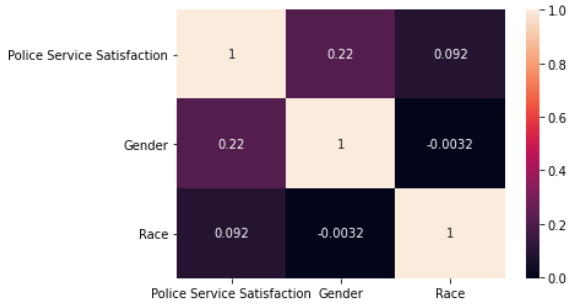

Figure 7 A heatmap of Race, Gender, and Police Satisfaction

Looking at the many measures of fitness, our model is undoubtedly less than ideal. This could be attributed to a number of reasons. To begin, it could simply be the case that race and gender were not such influential features ten years ago as they are today. It could also be possible that using race and gender are too few features to calculate an accurate prediction of police service satisfaction. It is possible that Logistic Regression was not ideal for the dataset among the numerous modeling techniques. In closing, this experiment offered valuable information that race and gender do not produce satisfactory results when predicting police service satisfaction.

5. References

Bolger, Michelle A, et al. "What Matters in Citizen Satisfaction with Police: A Meta-Analysis." *Journal of Criminal Justice*, Pergamon, 9 Nov. 2020, https://www.sciencedirect.com/science/article/pii/S0047235220302543.

Ng, Andrew Y, and Michael I Jordan. *On Discriminative vs. Generative Classifiers: A ... - Neurips*. University of California, Berkley, https://proceedings.neurips.cc/paper/2001/file/7b7a53e239400a13bd6be6c91c4f6c4e-Paper.pdf.

"2012 Constitution Satisfaction Survey Results." *2012 Constitution Satisfaction Survey Results - CKAN*, Arlington County, 12 Nov. 2020, https://catalog.data.gov/dataset/2012-constitution-satisfaction-survey-results.