Perceptions of Security in Mexico

HarvardX PH125.9x Capstone

Octavio Rodríguez Ferreira

6/10/2021

Contents

1 Executive summary

This project intends to determine whether trends on crime and violence are the only factors influencing perceptions of security, or if other more nuanced factors also have an important wight in shaping such perceptions. We argue that perceptions of security, while certainly determined by "hard" data such as crime, victimization and violence; are indeed influenced by other factors, and that they have a different impact across segments of population. Far from trying to find those "unknown" more nuanced factors, this project aims to prove how those factors influence different groups. Through different machine learning analysis, we predict almost 70% of responses to a Mexican public security perceptions survey, only testing

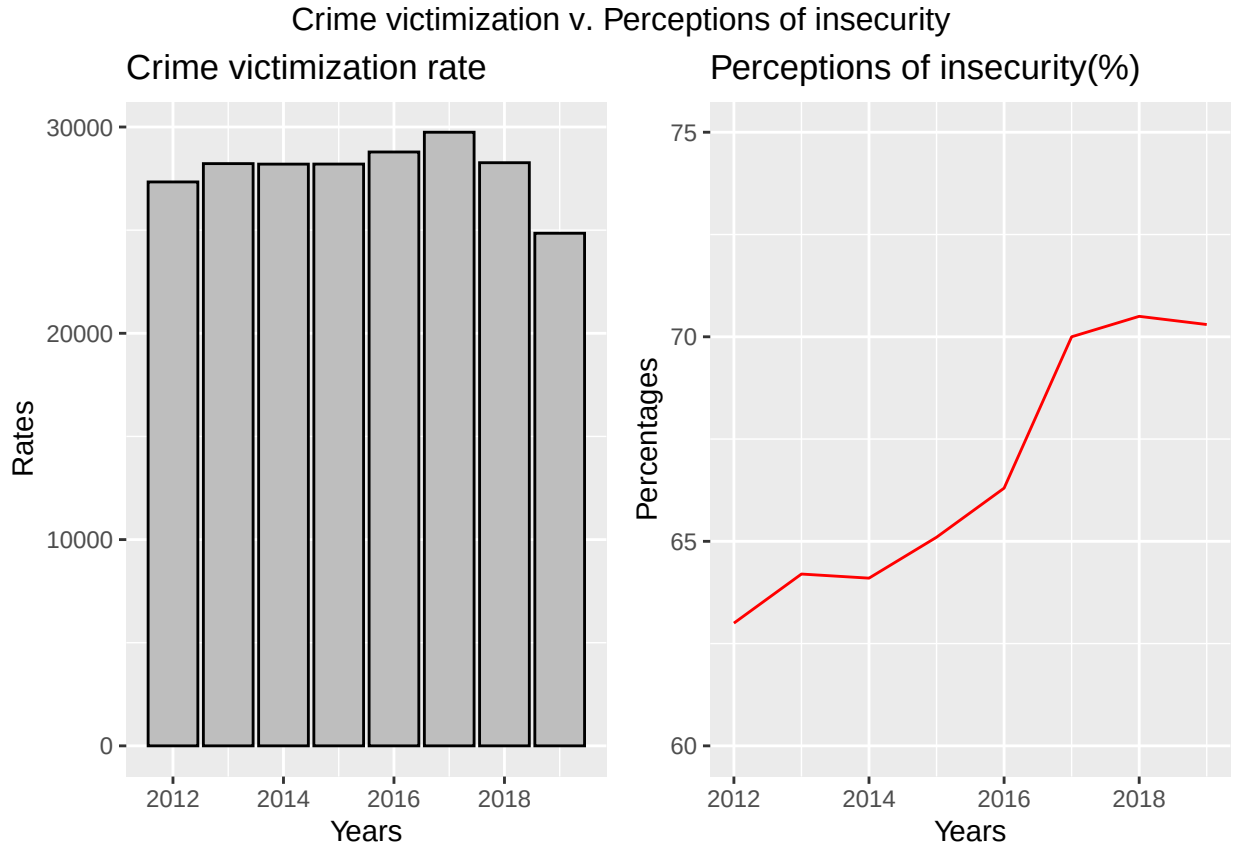
certain variables such as gender, age, type of area of residence (rural v. urban), SES, and the municipality of residence. Thus we conclude that trends on crime and violence are not the only factors influencing perceptions of security, and that there are other factors that influence differently various population groups.

2 Introduction

Mexico is considered one of the most violent countries in the world. Over the last fifteen years, there has been a steady increase in intentional homicides and other violent crimes. Violence in Mexico has have record-high numbers since 2017 and a steady increase since 2015, this measured by intentional homicides, which is a proxy to measure violent crime and a good indicator of levels of security (UNODC 2014) and "instrumental violence" (van Dijk 2008, 157). In 2020, amid the COVID-19 pandemic, crime and murder rates appeared to have leveled off, but remained at historically high levels.

The toll that the increase in violence and in violent crimes has taken on Mexican society has been evident. Since 2015, public's negative perception of security [hereinafter POS] has been a constant. According to several surveys by Mexico's National Institute for Statistics and Geography (Instituto Nacional de Estadística y Geografía, INEGI), the percentage of Mexicans who feel safe living in their respective municipality or city is only about 30% with minor fluctuations year by year, which means that around two thirds of the Mexican population feel unsafe living in their municipality or city (Justice in Mexico, 2016).

However, while perceptions of insecurity in Mexican municipalities have increased from 63% in 2013 to 70% in 2019, the rate of crime victimization has stayed fairly stable, and even decreased in 2018 and 2019, according to INEGI's National Survey of Crime, Victimization and Public Security Perceptions (Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública, ENVIPE).



So the question arises if the POS is exclusively influenced by crime and violence rates, or if there other

factors that help to shape own perceptions of insecurity. In other words, can one feel unsafe without having being a victim of crime? or, better yet, if increases in perception of insecurity necessarily reflect increases in actual victimization.

Our hypothesis is, no. We believe there are other more nuanced factors that have an important weight in shaping such perceptions. We do not intend, however, to find those factors, but to determine how certain "hidden" causes are relevant in own perceptions of security, and how those causes affect differently, different population groups. Overall, we argue that only by considering a handful of sociodemographic variables, we can very accurately predict perceptions of security and insecurity, which would be an indication of the existence of those "hidden" factors that influence such perceptions.

In order to achieve this, we use data from INEGI's ENVIPE survey from 2020, where we use variables such as age, gender, SES, municipality and the type of area where the respondents live (rural, urban or suburban) and a response to a question on whether respondents perceive their municipality as secure or insecure.

We use machine learning algorithms such as logistic regression, naive bayes and random forest to try to predict responses on POS using the sociodemographic variables mentioned above. With our most successful model, we can predict almost 70% of such responses.

We acknowledge that an accuracy of 67% is not as strong statistically, but it does help to demonstrate our hypothesis, since our model was able to predict respondent's perceptions in the majority of cases (2 out of 3 times), by testing simple sociodemographic indicators, and without without incorporating other variables that could be more influential in such a measure (crime and victimization rates, etc.).

Thus we conclude that our model is solid enough to demonstrate that trends on crime and violence are not the only factors influencing perceptions of security, and that we can infer the existence of other "hidden" determinants that relate to different characteristics of the respondent as part of a social group.

3 Initial Exploration

3.1 Data features

The original ENVIPE data set has several problems that had to be addressed. It has several columns, most of which we won't use, the column names are illegible, there are blank spaces within values, all columns are of a `character` class, and "No Response" was coded with the number `9`. So we took following steps to clean the data:

- Remove all blank spaces
- Delete unnecessary columns
- Change classes
- Create a unique number for the municipality of respondent
- Recode the existing variables
- Remove the "No Response" coded as "9" from our dependent variable (POS)
- Add a new variable for "age group" to see if it performs better than the "age" variable that is a continuous value.

With all the changes described above, we end up with a data set of `89171` observations and `8` variables.

	Class	Variable description	Values
id	character	Id of respondent	Unique number
sex	numeric	Sex of respondent	Male = 1, Female =2
age	numeric	Age of respondent	11 to 80
age_grp	factor	Age group of respondent	-20, 20–39, 40–59, 60-79, 80+
ses	numeric	Socio-economic strata	Low = 1, Medium-low = 2, Medium-high = 3, High = 4
area	factor	Type of area of residence	Urban = 1, Suburban = 2, Rural = 3
mun	numeric	Unique code for municipality of residence	Code composed of State and Municipality official numbers
pos	factor	Perception of security in the municipality	1 = Secure, 2 = Insecure

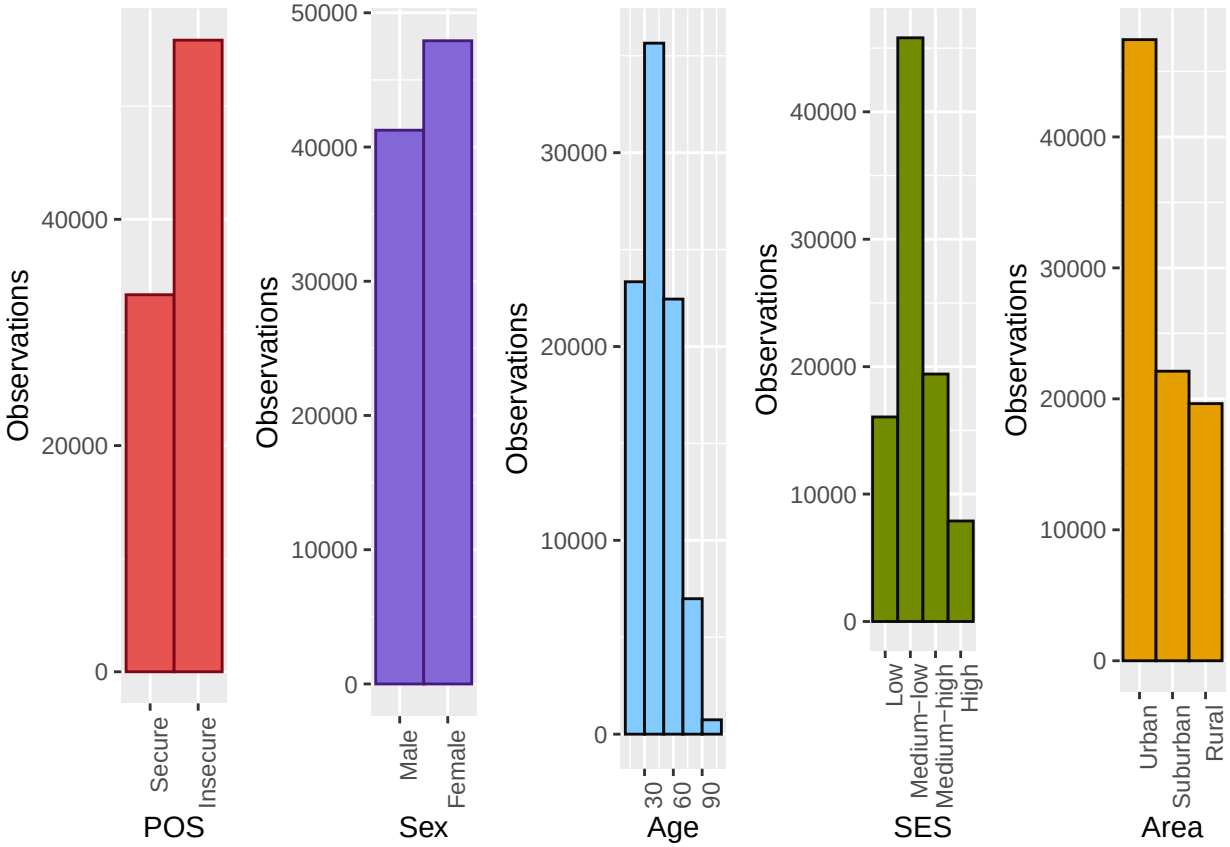
3.2 Descriptive Analysis

The distribution of respondents to the ENVIPE survey is fairly well distributed geographically. That means we can have perspectives from people in different contexts and with different backgrounds.

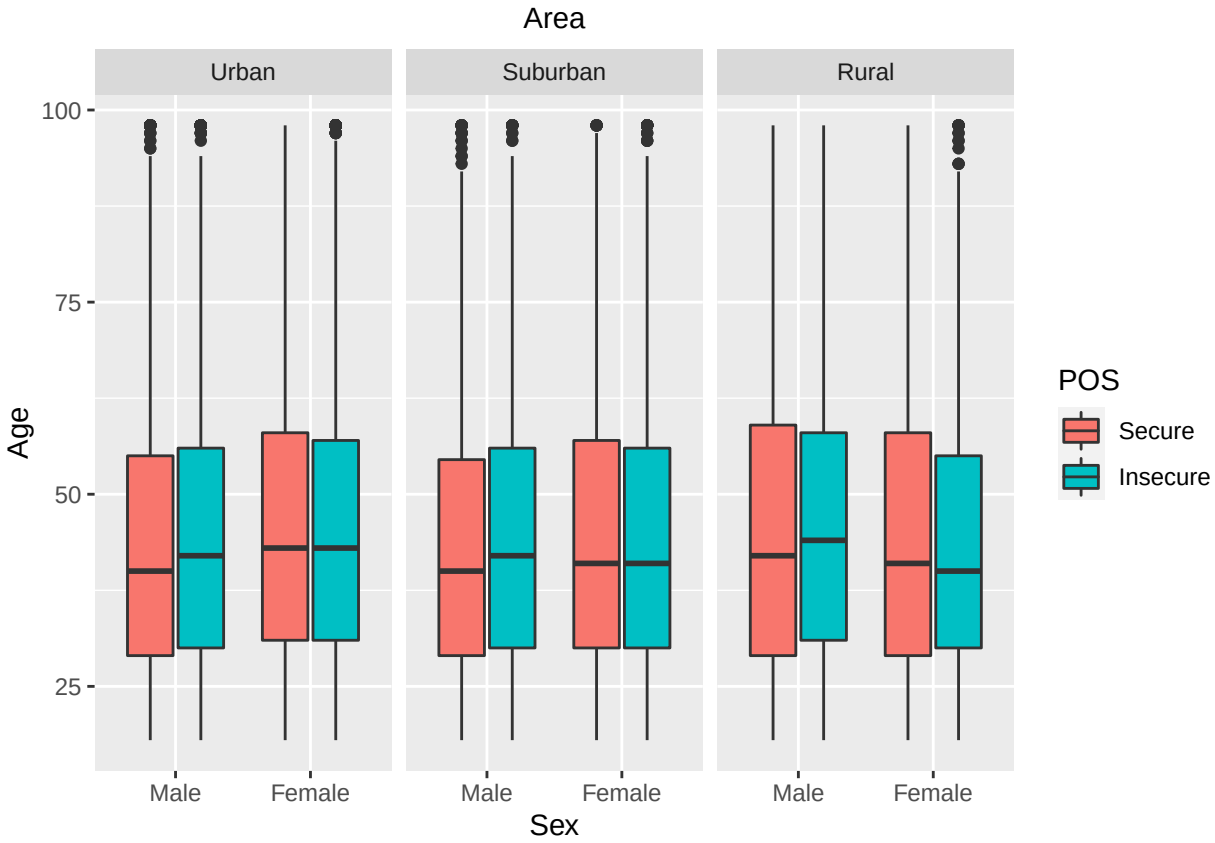
```
##
##      checking for file ‘/private/var/folders/fd/x4blh2qd0s31f6fs0jhp9cgxsqmjff/T/RtmpzfCvBH/remotese
##   -  preparing ‘mxmaps’:
##      checking DESCRIPTION meta-information ...  v  checking DESCRIPTION meta-information
##   -  checking for LF line-endings in source and make files and shell scripts
##   -  checking for empty or unneeded directories
##   -  building ‘mxmaps_2020.0.0.tar.gz’
##      Warning: invalid uid value replaced by that for user 'nobody'
##    Warning: invalid gid value replaced by that for user 'nobody'
##
##
```

Geolocation of respondents

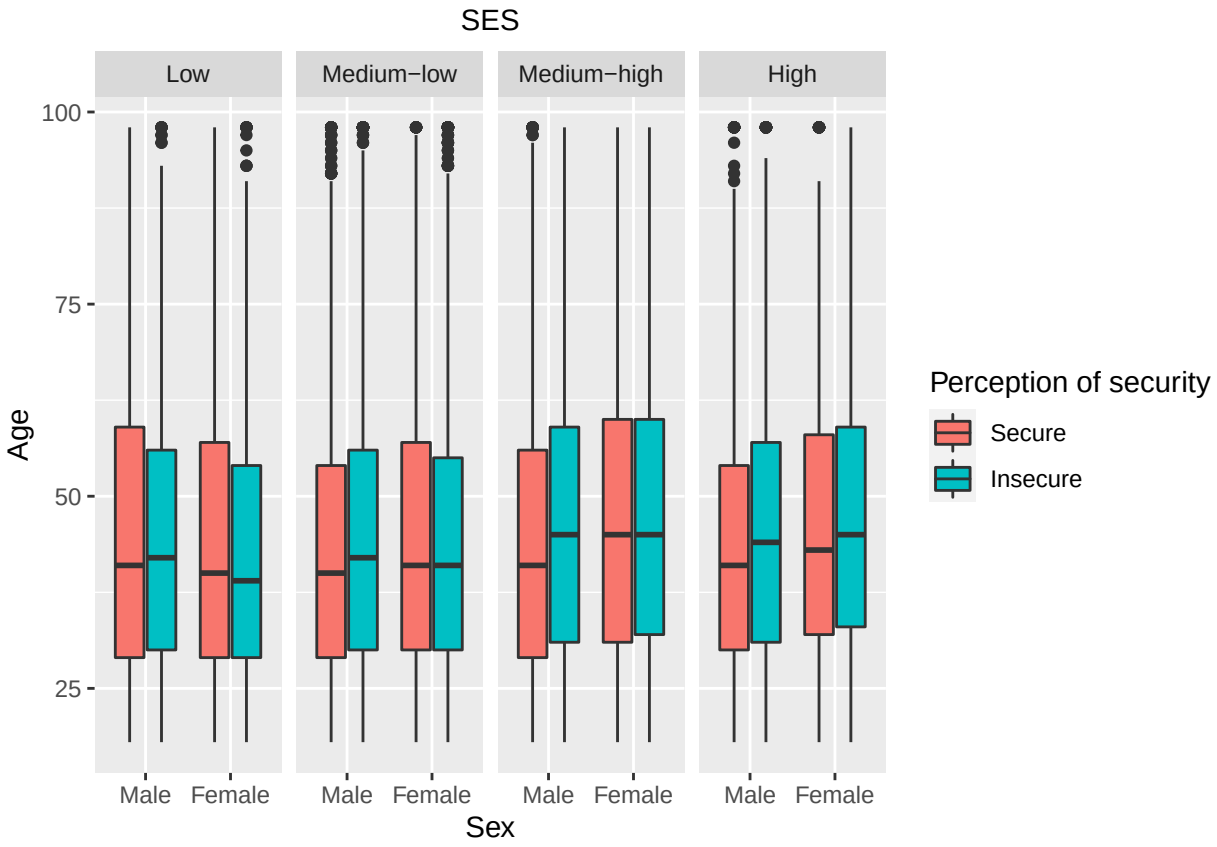
From an exploratory analysis of the data, we found that almost twice as much of the respondents feel their municipality to be insecure, and women tend to feel less secure than men. In terms of the sociodemographic characteristics, the majority of respondents is younger than 50 years old, mostly between 30 and 50 years old. Also, the vast majority of respondents are middle class (low-middle and high-middle), though the distribution is skewed toward medium-lower and lower class. Finally, the vast majority of respondents come from a urban area and those who come from a sub-urban and rural areas are almost equally distributed.



When combined, we can see how different variables interact with each other. For example, on average people from urban areas tend to feel more secure, however younger women from rural areas tend to feel more insecure while older men tend to feel more secure. In urban areas men also feel safer than women, but older men tend to feel more insecure than younger men. In urban areas, the distribution of women who feel either secure or insecure by age is fairly similar. In suburban areas we see almost the same patterns as in urban areas.



In regards on how the variables distribute based on SES, we have that the lower class tends to feel more secure as higher classes. Also in lower classes, younger women tend to feel more insecure that men, however those women who say that they feel secure are very balanced with the ones that feel insecure. In medium classes we see a more diverse distribution than with high and low classes. For example, the distribution between those who women who feel secure and insecure is almost the same while older men from the same social class tend to feel more insecure. Though, We see very similar patterns in the both strata of the middle class.



4 Methods and analysis

Before starting the modeling stage, the data needed to be partitioned. To perform a machine learning analysis, it is necessary to divide our data set in two different sets, one for training our algorithms and one for testing.

4.1 Data partitions

The initial partition of our data consisted of 80% of the original observations as our `training` set, and the `validation` set the remaining 20%, to test the final model. We made an additional partition splitting in half the `training` set into a `train_set` and a `test_set`, to train and test different models much faster and easier.

4.2 Modeling

The first step of our modeling stage was to determine the type of analysis and variables to use. In this stage we used logistic regression model and tested different variables to determine the best combination of predictors.

The baseline model was then defined as a logistic regression analysis that included the variables `sex`, `age`, `mun`, `ses`, and `area` as predictors, with an initial accuracy of roughly `0.628`.

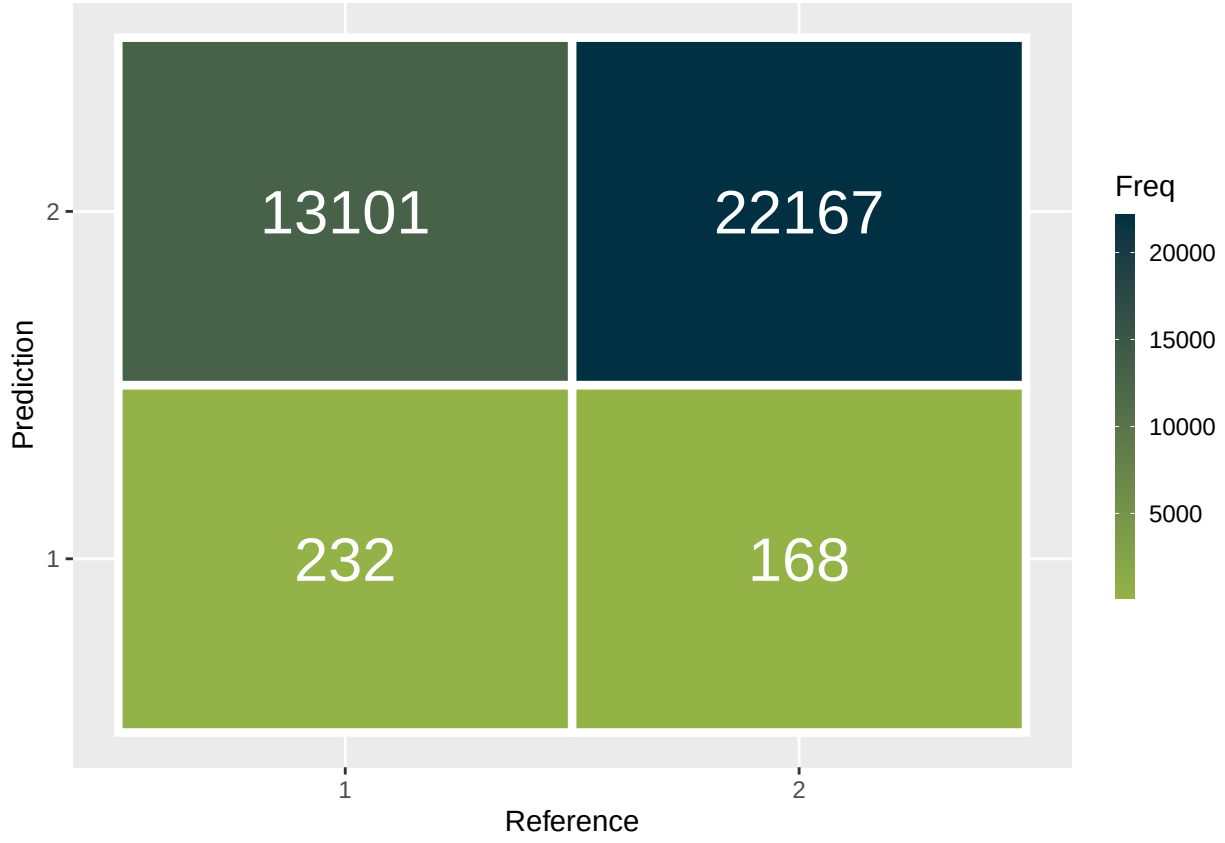
Predictors	Accuracy
Sex+Age+Municipality	0.6261915
Sex+Age+Municipality+SES	0.6261915
Sex+Age+Municipality+Area	0.6261915
Sex+Age+Municipality+SES+Area	0.6279859

Once defined the variables and the logistic regression as our baseline analysis we tried also a naive bayes and a random forests analysis. We also alternated the use of the `age` and `age_grp` variables with two of the models to see which one yielded a better accuracy.

4.2.1 Logistic Regressions

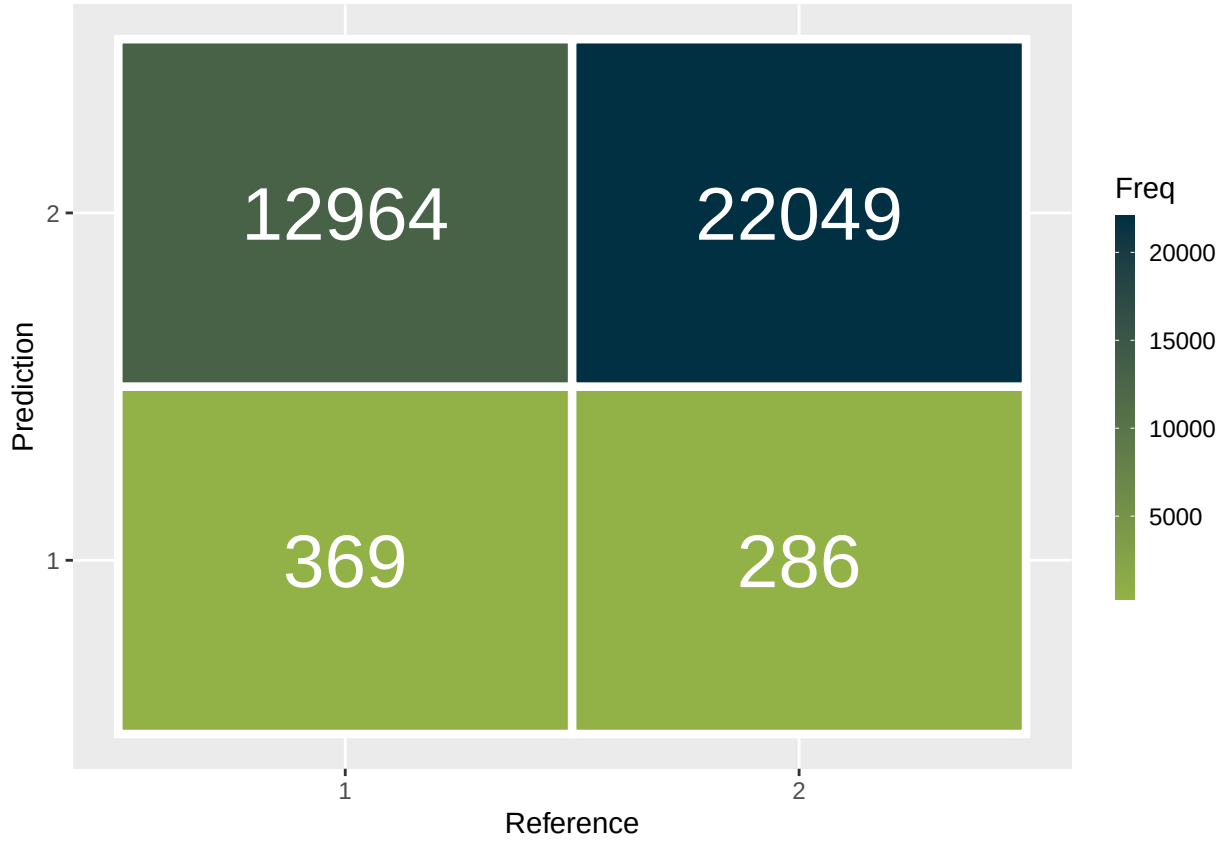
In the first logistic regression model we used the `age` variable obtaining an accuracy of almost `0.628`.

Accuracy	Kappa	AccuracyLower	AccuracyUpper	AccuracyNull	AccuracyPValue	McnemarPValue
0.6279859	0.0122788	0.6229455	0.6330054	0.6261915	0.2436168	0



We then tried an alternative version of the same baseline model using `age_grp` instead of `age`. The adjusted logistic regression model yielded an accuracy of `0.6285`, which was improvement from the the previous model.

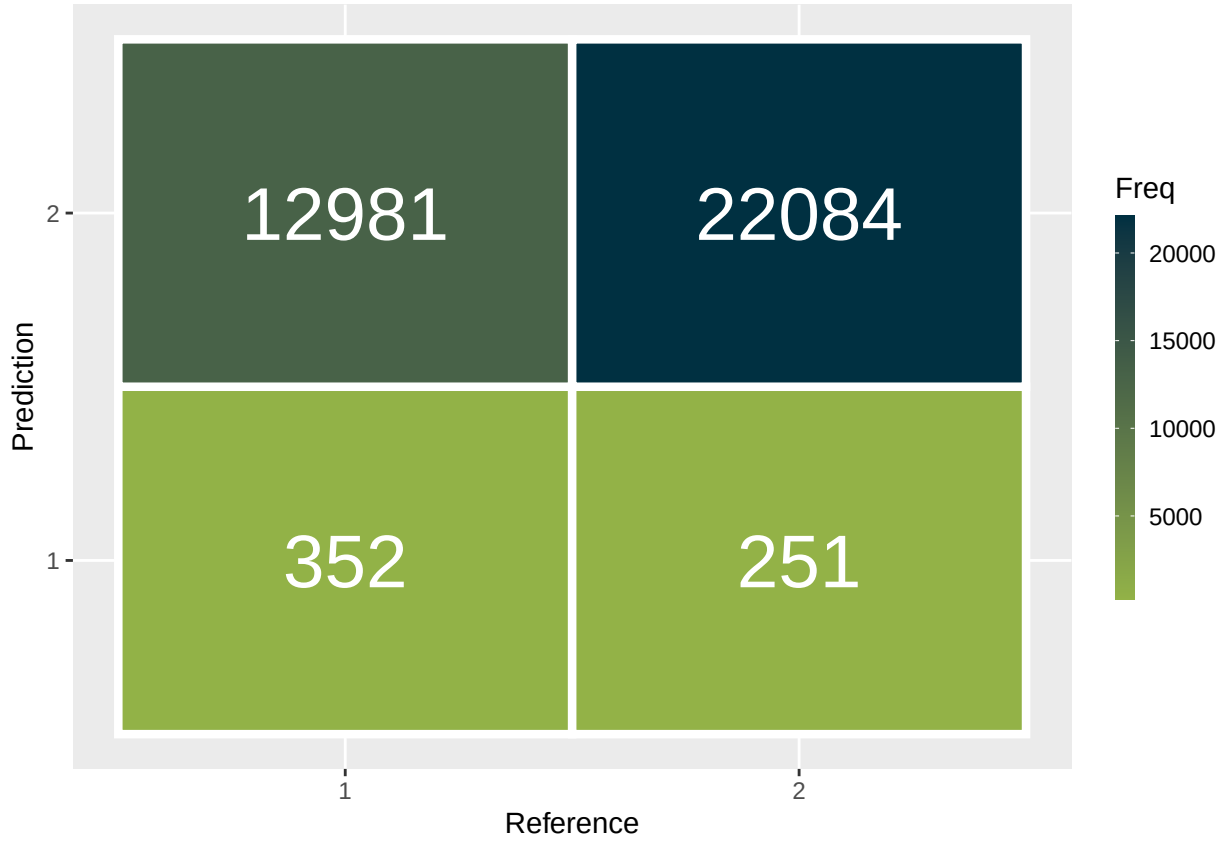
Accuracy	Kappa	AccuracyLower	AccuracyUpper	AccuracyNull	AccuracyPValue	McnemarPValue
0.6285186	0.0183957	0.6234796	0.6335365	0.6261915	0.1833139	0



4.2.2 Naive Bayes

The Naive Bayes algorithm proved to be more accurate than the logistic regression. When trained with `age` variable it gave almost `0.63`. With Naive Bayes we didn't test `age_grp` because the type of variable was not compatible with the algorithm.

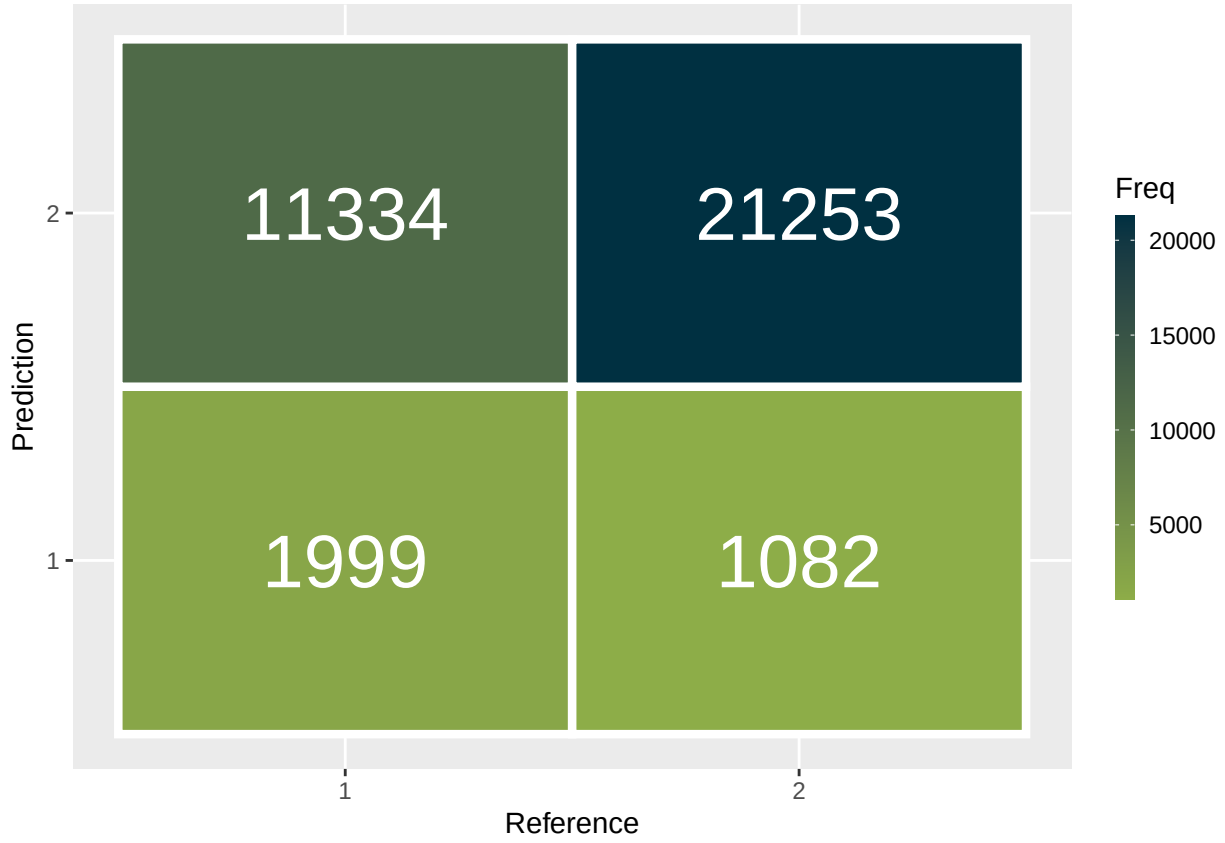
Accuracy	Kappa	AccuracyLower	AccuracyUpper	AccuracyNull	AccuracyPValue	McnemarPValue
0.6290232	0.0187752	0.6239856	0.6340398	0.6261915	0.1356692	0



4.2.3 Random Forests

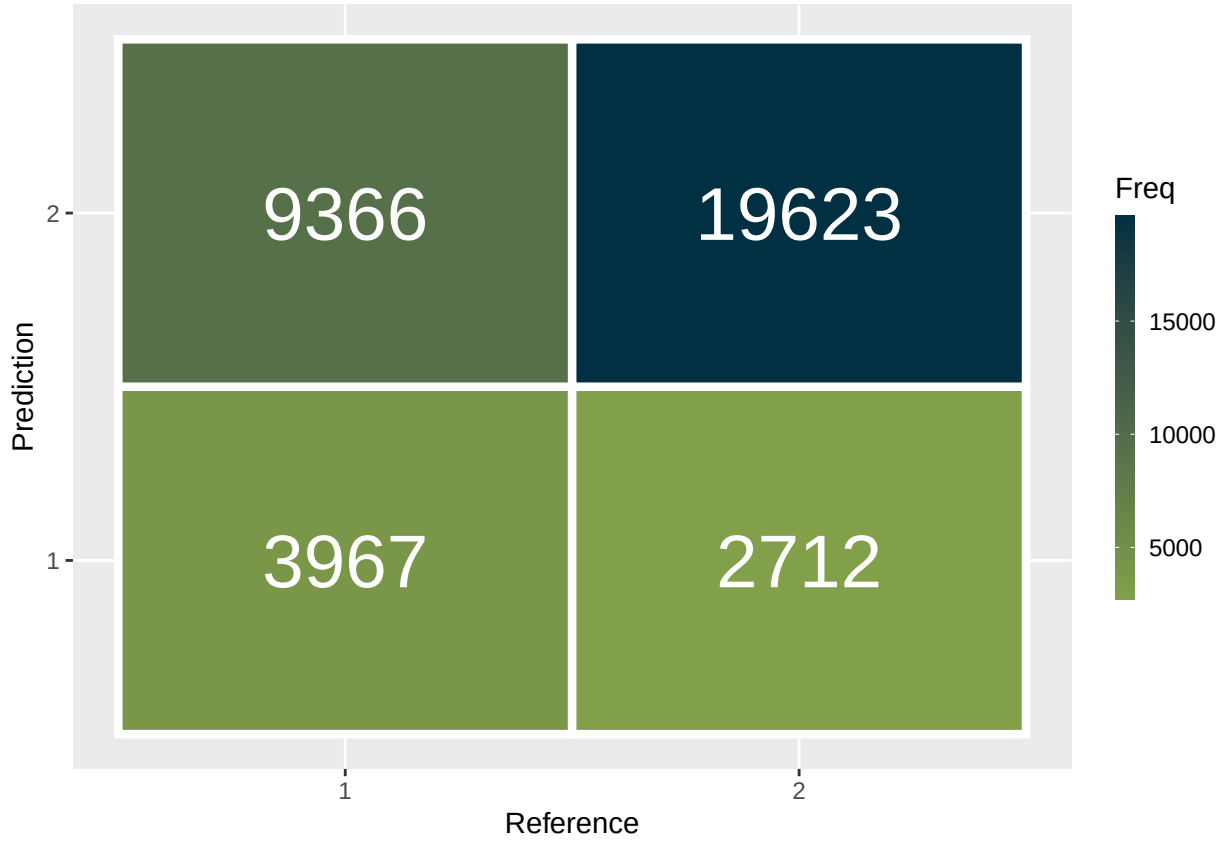
The Random Forests model proved to be the most effective algorithm in the training stage increasing our accuracy to roughly `0.652` using the `age` variable.

Accuracy	Kappa	AccuracyLower	AccuracyUpper	AccuracyNull	AccuracyPValue	McnemarPValue
0.6519009	0.1200936	0.6469309	0.656846	0.6261915	0	0



The algorithm performed even better with the `age_grp` variable, giving us an accuracy of a little over `066`.

Accuracy	Kappa	AccuracyLower	AccuracyUpper	AccuracyNull	AccuracyPValue	McnemarPValue
0.6613771	0.1958008	0.6564389	0.6662891	0.6261915	0	0



Based on the accuracy increase with the adjusted Random Forests algorithm, we choose it as our final model to train and test in the larger `training` and `validation` sets.

Method	Accuracy
Baseline Logistic Regression	0.6279859
Adjusted Logistic Regression	0.6285186
Naive Bayes	0.6290232
Random Forests	0.6519009
Adjusted Random Forests	0.6613771

5 Results

The final Random Forests algorithm trained on the `training` set and tested on the `validation` set incorporated the `sex`, `age_grp`, `mun`, `ses`, and `area` variables as predictors. The code for the algorithm for the R language

```
##########################################################
#############      RANDOM FORESTS  MODEL     #############
##########################################################

# Model of the entire training set tested in validation.
# With Age Group instead of age.

#Train model
```

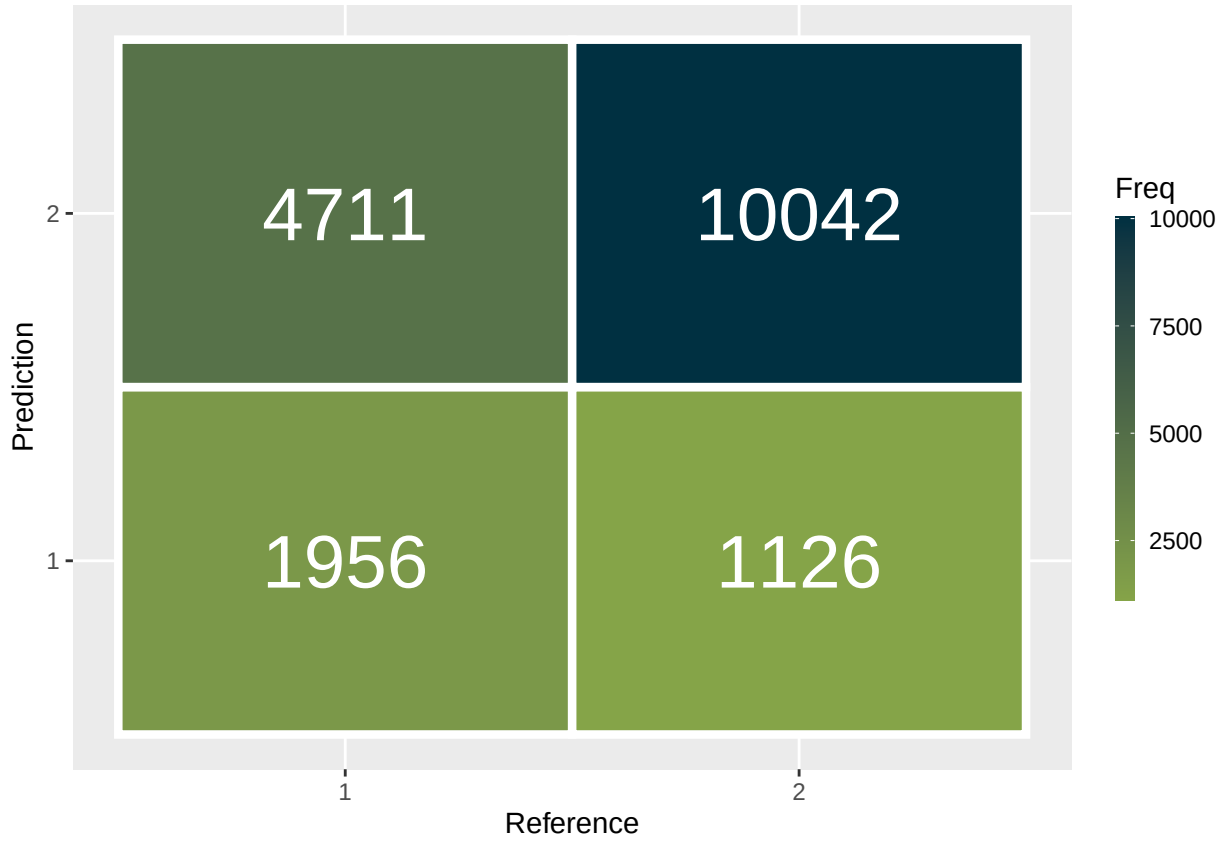
```r
fit_final <- train(pos ~
                     sex +
                     age_grp +
                     mun +
                     ses +
                     area,
                   method = "rf",
                   data = training)

#Predict
pred_final <- predict(fit_final, validation)

#Confusion matrix
cm_final <- confusionMatrix(data = pred_final, reference = validation$pos)
```

This model gave us a final accuracy of `0.6727` which meant an increase of `4.5%` over the original baseline model.

Accuracy	Kappa	AccuracyLower	AccuracyUpper	AccuracyNull	AccuracyPValue	McnemarPValue
0.6727222	0.2159629	0.6657801	0.6796079	0.6261845	0	0



It appears that sociodemographic variable might have a limited power as sole predictors of a person's opinion on a given issue. However with a convination of variables and adjustments we were able to predict a participant's opinion accurately almost `70%` of the times, which we believe is enough to partially demonstrate

our hypothesis.

Method	Accuracy
Random Forests	0.6727222

6 Conclusion

The increase in violence and the high levels of crime in Mexico certainly have changed the perceptions of security among its citizens. However, it appears that the popular perceptions of security can be influenced by many other factors.

Perceptions of security, while certainly determined by crime, victimization and violence; are indeed influenced by other factors, and that they have a different impact across segments of population. As stated above, far from trying to find those "hidden" more nuanced factors, we found that by simply using some sociodemographic variables we could accurately predict the POS of a respondent in almost 70% of the times, even without considering variables of crime victimization or crime rates.

This, of course doesn't mean that those external factors does not influence POS. On the contrary, those are the main drivers. However, what we try to demonstrate is that despite such factors, perceptions follow a pattern that might also be strongly influenced by other social factors.

7 References

UNODC. 2014. Global Study on Homicide 2013. United Nations Office on Drugs and Crime. Vienna: United Nations.

Van Dijk, Jan. 2008. The World of Crime. Thousand Oaks: Sage.

INEGI. 2020. Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública. https://www.inegi.org.mx/contenidos/programas/envipe/2020/datosabiertos/conjunto_de_datos_envipe2020_csv.zip

Justice in Mexico. 2016. Public's perception of security in Mexico stays same despite rise in homicides. San Diego: Justice in Mexico. https://justiceinmexico.org/publics-perception-of-security-in-mexico-stays-same-despite-rise-in-homicides/