# Violent or Non-violent

**Max Ouellette, Sally Alakeel, Bao Khoi Bui, Luis Waldo, and Gabriel Mena**

## Introduction:

**Sally Alakeel**

Crime in the United States has been on a steady decline since the 1990's however in some cities crime remains a major problem many citizens deal with on a daily basis. Being able to understand important crime statistics may allow for decision makers to identify and mitigate risk factors in their localities.

The goal of this project is to use and process data to answer the following questions: Given identifying statistics of an individual are they more likely to be the victim of a crime and if so is that crime non-violent or violent in nature using a Logistic Regression and Random Forest models alongside cross validation to determine which of these models best fits the data.

### About The Data:

For our research we found a dataset on the Data.gov catalog related to crimes committed in Los Angeles between 2010 and 2019. We predict that with the knowledge of certain attributes of a specific crime instance we can then use that information with other predictors to determine whether a person is at high risk of being involved in a violent or non-violent crime instance. This data set is sufficiently large coming in at 2.13M observations over 28 variables. For each observation there are 28 variables, however some are removed due to being redundant. The most important variables are the ID# of the crime, the dates and times, area, type of crime, statistics of the victim, weapon (if any), and location.

### The Models:

We chose to use Logistic Regression and Random Forest models for this report. The individual sections for both are detailed below.

## Logistic Regression Model

**Bao Khoi Bui, Max Ouellette**

The objective of this model is to predict the likelihood that an individual will be a victim of a violent or non-violent crime based on the location, age, race, and other personal details. The reason for a logistic regression model is to determine which crimes are violent and nonviolent comes down to a classification problem, where our response variable (categorical) is if the crime

is violent, and the predictors will be the factors that the response will be dependent on, meaning any change in any one of the predictors will influence a change in the response as well. The advantages of logistic regression are:

1. Allows us to interpret the coefficients of our model, seeing how each predictor impacts the response in what way.
2. Computationally efficient for small to medium size data sets.

The disadvantages of logistic regression are:

1. It is not very good with complex or non-linear predictors. In other words, the model may not be able to capture the relationship between the response and those types of predictors without some kind of intervention that makes the predictors more suitable to the model.
2. It is sensitive to multicollinearity, which could affect how our coefficients will be interpreted if they are highly correlated with each other.

Below is the equation to our logistic regression model, which is shown as:

$$P(y = isViolentCrime|x) = \frac{exp(\beta_0 + \beta_1*Age + \beta_2*Time + \beta_3*isWhite + \beta_4*isBlack + \beta_5*isInAlley + \beta_6*isInApartment}{1 + exp(\beta_0 + \beta_1*Age + \beta_2*Time + \beta_3*isWhite + \beta_4*isBlack + \beta_5*isInAlley + \beta_6*isInApartment}$$

$$\frac{+\beta_7*Business + \beta_8*isInEducationLocation + \beta_9*isInGovernmentFac + \beta_{10}*isInHealthcareFac + \beta_{11}*isInHotel + \beta_{12}*isInHouse}{+\beta_7*Business + \beta_8*isInEducationLocation + \beta_9*isInGovernmentFac + \beta_{10}*isInHealthcareFac + \beta_{11}*isInHotel + \beta_{12}*isInHouse}$$

$$\frac{+\beta_{13}*isInJail + \beta_{14}*isInMotel + \beta_{15}*isInOtherPrem + \beta_{16}*isInOtherHome + \beta_{17}*isInRoad + \beta_{18}*isInTownHouse + \beta_{19}*isInWork}{+\beta_{13}*isInJail + \beta_{14}*isInMotel + \beta_{15}*isInOtherPrem + \beta_{16}*isInOtherHome + \beta_{17}*isInRoad + \beta_{18}*isInTownHouse + \beta_{19}*isInWork}$$

$$\frac{+\beta_{20}*isHispanic + \beta_{21}*LAT + \beta_{22}*LON)}{+\beta_{20}*isHispanic + \beta_{21}*LAT + \beta_{22}*LON)}$$

As shown, our logistic regression equation contains 22 predictors that will help in predicting the likelihood of a crime committed to the victim is violent or non-violent. Since we are working with a large dataset, we will limit the random samples to 2,500 for better efficiency.

```
crimeDataFull = subset(crimeDataFull, Vict.Age != 0)
crimeDataFull = subset(crimeDataFull, crimeDataFull$Crm.Cd != 944)
crimeDataFull = subset(crimeDataFull, crimeDataFull$Crm.Cd != 954)
crimeDataFull = subset(crimeDataFull, crimeDataFull$Premis.Cd != 601)
crimeDataFull = subset(crimeDataFull, crimeDataFull$Premis.Cd != 750)
crimeDataFull = subset(crimeDataFull, crimeDataFull$Premis.Cd != 803)
crimeData = crimeDataFull[sample(1:nrow(crimeDataFull), 2500), ]
```

We will exclude some features and values of certain features from our dataset due to its irrelevance to the objective of our model. The value being excluded: victim's age is 0. The features being excluded: conspiracy & contributing crimes and savings & loan, cyberspace, and retired. Below is the summary of our model, showing us the coefficients and significant values

for each of the predictors. Based on the model's p-values, we can see that most of the model's predictors are significant with the only exceptions that are not significant being Time.OCC, isWhite, isInGovernmentFac, isInHotel, isInJail, isInWork, LAT, and LON. This means 14 out of the 22 features in the model are statistically significant in predicting whether an individual will be a victim of a violent crime.

```
summary(crimeData_pred)
```

```
Call:
glm(formula = isViolentCrime ~ Vict.Age + TIME.OCC + isWhite +
    isBlack + isInAlley + isInApartment + isInBusiness + isInEducationLocation +
    isInGovernmentFac + isInHealthcareFac + isInHotel + isInHouse +
    isInJail + isInMotel + isInOtherPrem + isInOtherHome + isInRoad +
    isInTownhouse + isInWork + isHispanic + LAT + LON, family = "binomial",
    data = crimeData)


Coefficients: (1 not defined because of singularities)
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -5.740e-01  1.456e+00  -0.394 0.693462
Vict.Age             -2.369e-02  3.146e-03  -7.528 5.14e-14 ***
TIME.OCC              1.665e-05  6.977e-05   0.239 0.811347
isWhite               1.245e-01  1.654e-01   0.753 0.451455
isBlack               9.106e-01  1.709e-01   5.328 9.93e-08 ***
isInAlley             1.305e+00  4.882e-01   2.673 0.007523 **
isInApartment         1.501e+00  2.548e-01   5.891 3.85e-09 ***
isInBusiness          5.600e-01  2.731e-01   2.050 0.040329 *
isInEducationLocation 1.250e+00  3.775e-01   3.312 0.000927 ***
isInGovernmentFac     1.388e+00  8.162e-01   1.701 0.089017 .
isInHealthcareFac     2.331e+00  8.238e-01   2.830 0.004661 **
isInHotel             9.113e-01  9.002e-01   1.012 0.311357
isInHouse             1.009e+00  2.501e-01   4.033 5.52e-05 ***
isInJail                     NA         NA      NA       NA
isInMotel             1.607e+00  7.520e-01   2.136 0.032656 *
isInOtherPrem         6.045e-01  2.752e-01   2.197 0.028054 *
isInOtherHome         7.695e-01  3.698e-01   2.081 0.037429 *
isInRoad              1.505e+00  2.483e-01   6.062 1.34e-09 ***
isInTownhouse         2.172e+00  1.089e+00   1.995 0.046022 *
isInWork              1.086e-02  2.754e-01   0.039 0.968557
isHispanic            6.455e-01  1.542e-01   4.185 2.85e-05 ***
LAT                  -6.898e-01  4.416e-01  -1.562 0.118255
LON                  -1.913e-01  1.277e-01  -1.498 0.134178
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3156.5  on 2499  degrees of freedom
Residual deviance: 2882.9  on 2478  degrees of freedom
AIC: 2926.9
```

In order to look at the effectiveness of our model, we split our data into eighty percent training and twenty percent testing.

```
#Splitting things into training and testing data
set.seed(2000)
meanPredError = 0
for(i in 1:10){
  trainCutOff = sample(1:nrow(crimeData), nrow(crimeData)*0.8)
  crimeDataTrain = crimeData[trainCutOff,]
  crimeDataTest = crimeData[-trainCutOff,]
  crimeDataTrainPred = glm(isViolentCrime ~ Vict.Age + TIME.OCC + isWhite + isBlack + isInAlley +
                             isInApartment + isInBusiness + isInEducationLocation +
                             isInGovernmentFac + isInHealthcareFac + isInHotel + isInHouse +
                             isInJail + isInMotel + isInOtherPrem + isInOtherHome + isInRoad +
                             isInTownhouse + isInWork + isHispanic + LAT + LON,
                           data = crimeDataTrain, family="binomial")
  crimeDataTestPred = glm(isViolentCrime ~ Vict.Age + TIME.OCC + isWhite + isBlack + isInAlley +
                            isInApartment + isInBusiness + isInEducationLocation +
                            isInGovernmentFac + isInHealthcareFac + isInHotel + isInHouse +
                            isInJail + isInMotel + isInOtherPrem + isInOtherHome + isInRoad +
                            isInTownhouse + isInWork + isHispanic + LAT + LON,
                          data = crimeDataTest, family="binomial")
  crimeDataMSE = mean((crimeDataTest$isViolentCrime -
                         predict(crimeDataTrainPred, newdata=crimeDataTest,type="response"))^2)
  meanPredError = meanPredError + sqrt(crimeDataMSE)
  print(sqrt(crimeDataMSE))
}
meanPredError = meanPredError / 10
print(meanPredError)
```

```
[1] 0.4459377
[1] 0.4413151
[1] 0.4367835
[1] 0.4534802
[1] 0.4447312
[1] 0.462509
[1] 0.4402415
[1] 0.4508206
[1] 0.4585523
[1] 0.4372822
> meanPredError = meanPredError / 10
> print(meanPredError)
[1] 0.4471653
```

After splitting our data into training and testing, all of the features used to predict the likelihood of whether an individual will be a victim of a violent or non-violent crime are still highly significant. The reason for this is due to the large amount of data points that our data set has. As a result, the model can capture the effects of some predictors that might be negligible but are still statistically significant. After ten iterations of splitting, we see that we have the lowest MSE being 43.67835% and the highest being 46.2509%. We then calculated the average of those ten MSE iterations to get our mean prediction error of 44.71653%. This means that on average, our model's predictions are incorrect about 44.71653% of the time, which if our error threshold was 50%, we could say that our model is performing a little better than just randomly guessing. So when it comes to predicting if a crime is violent or not, the model may not be the greatest as

there are still improvements to it that can be made such as whether to increase its performance or lower its prediction error, but it can still be used for our objective.

## Random Forest Model

**Gabriel Mena, Luis Waldo**

A random forest model is advantageous for this analysis because it is well-suited for classification tasks involving a mix of categorical and numerical variables, such as those in the L.A. crimes dataset. Random forests are ensemble models that construct multiple decision trees during training and output the mode of the classes for classification tasks. This approach provides several advantages for the study.

Random forests aggregate multiple decision trees, improving accuracy and reducing the risk of overfitting. They effectively capture complex relationships between predictors like TIME.OCC, AREA, Vict.Age, Vict.Sex, Vict.Descent, and Premis.Cd. Additionally, the model provides valuable insights into variable importance, helping to identify key factors influencing crime classification.

However, random forests have some limitations. The model's complexity can make it difficult to interpret individual predictions. Training can also be computationally intensive, especially with a large number of trees, requiring more time and resources. Furthermore, the model's performance depends on careful tuning of hyper-parameters, such as the number of trees and features considered at each split, which can be a time-consuming process. Despite these drawbacks, the random forest's strengths in handling complex data and delivering accurate predictions make it a suitable choice for this analysis.

For the random forest model used in this analysis, the model formula can be written as:

$$\text{violent} \sim \text{TIME.OCC} + \text{AREA} + \text{Vict.Age} + \text{Vict.Sex} + \text{Vict.Descent} + \text{Premis.Cd}$$

This formula indicates that the response variable, *violent* (indicating whether a crime is violent or non-violent) is modeled based on the predictor variables: TIME.OCC (crime time of occurrence), AREA (LAPD geographic area codes), Vict.Age (victim's age), Vict.Sex (victim's sex), Vict.Descent (victim's ethnic and racial descent), and Premis.Cd (premises code). In the context of random forests, this formula is used to guide the splitting criteria for each decision tree in the ensemble, with the overall prediction being determined by aggregating the outputs of all trees.

In fitting the random forest model for this analysis, several considerations were taken into account to ensure meaningful insights. First, the inclusion of predictors such as TIME.OCC,

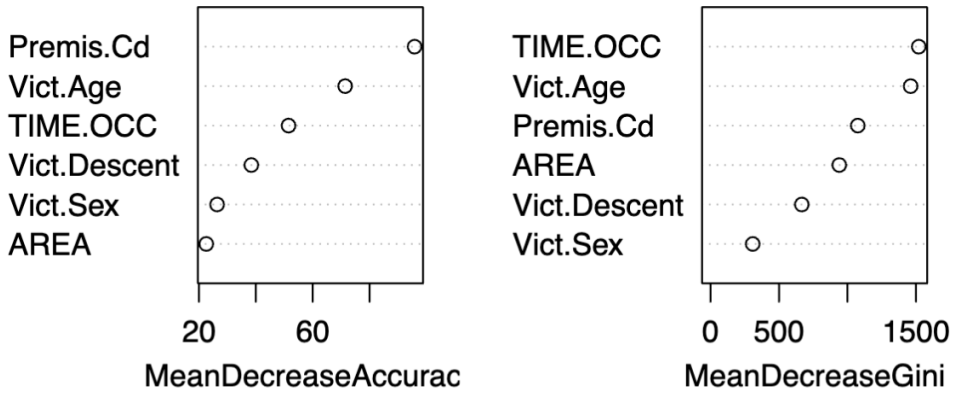
AREA, Vict.Age, Vict.Sex, Vict.Descent, and Premis.Cd was guided by both domain knowledge and their potential relevance to predicting violent crimes. Feature importance scores from the model will help determine if any predictors contribute minimally and can be excluded in future iterations. Hyper-parameter tuning: 500 trees as a balance between accuracy and computational efficiency. Additionally, mtry was set to 2 (following $mtry = \sqrt{p}$), ensuring that the model explores a subset of predictors at each split, promoting diversity among the trees and reducing the risk of overfitting

```
formula <- violent ~ TIME.OCC + AREA + Vict.Age + Vict.Sex + Vict.Descent + Premis.Cd
test_errors <- numeric(10)
for (i in 1:10) {
  set.seed(i)
  sample_index <- sample(seq_len(nrow(la_crime)), size = 0.8 * nrow(la_crime))
  train_data <- la_crime[sample_index, ]
  test_data <- la_crime[-sample_index, ]

  rf_model <- randomForest(formula, data = train_data,
                           ntree = 500, mtry = 2, importance = TRUE)

  predictions <- predict(rf_model, test_data)

  confusion_matrix <- table(test_data$violent, predictions)
  error_rate <- 1 - sum(diag(confusion_matrix)) / sum(confusion_matrix)

  test_errors[i] <- error_rate
}
print(paste("Mean Test Error:", round(mean(test_errors) * 100, 2), "%"))
```

```
[1] "Mean Test Error: 25.11 %"
```

The random forest model achieved a test error rate of 23-26%, indicating that it accurately classifies crimes as violent or non-violent in approximately 74-77% of cases. While this performance suggests the model captures meaningful patterns, there is still room for improvement. The error rate may stem from the inherent complexity of the dataset, potential noise in the predictors, or something more obscure. Despite these limitations, the model demonstrates strong predictive capability.

One of the key strengths of random forests is their ability to assess the importance of each predictor in the classification process. Below is the variable importance plot generated from the random forest model:




Noting that **MeanDecreaseAccuracy** measures how much model accuracy decreases when each variable is excluded and that **MeanDecreaseGini** measures the contribution of each variable to the homogeneity of the nodes and leaves in the trees. The following interpretations of each variable can be concluded:

- TIME.OCC appears as the most important variable in both metrics, suggesting that the time of occurrence of the crime significantly influences whether a crime is violent or non-violent.
- Vict.Age is consistently influential, indicating that the victim's age plays a critical role in predicting crime type.
- Premis.Cd has notable importance, signifying that the location context (type of premises) is critical.
- AREA and Vict.Descent also contribute, though to a slightly lesser extent.
- Vict.Sex appears least influential among these variables.

Based on the variable importance, time and location-specific factors are the strongest predictors of whether a crime will be violent. This aligns with the hypothesis that certain types of violent crimes are more likely to occur at specific times or in particular locations. Personal details of the victim, such as age and descent, also play a role but less than temporal and location-based factors. However, Victim's sex has limited predictive value for distinguishing violent from non-violent crimes in this dataset. This analysis suggests that efforts to prevent violent crimes could benefit from focusing on high-risk times and locations.

## Results

**Max Ouellette**

The logistic regression model had usage in showing us what predictors could be useful within seeing whether someone were to be a victim of a violent crime or not. The model showed us that when the crime occurred, certain premises which include government facilities, hotels, jails, work locations, latitude, and longitude are not important predictors when it comes to predicting whether a crime committed to you would be violent or not.

While the linear regression model gives an idea on what predictors should be used or not, inference is not what we are constructing our models for. We are constructing our models for the goal of prediction. When it comes to prediction, our linear regression model is not usable. The reason behind this is because the average error rate is 44.71653 percent, which is close to fifty percent. Because of the fact that our error rate is close to fifty percent, the reliability of our model is close to that of randomly guessing.

While the linear regression model is unreliable, the random forest model is much more reliable. The error rate is 23-26% meaning it can be used to predict whether a crime victim was a victim of a non-violent crime or not. Not only that, but it also shows which predictors are more important than others. According to the random forest model, we can see that premise, time occurred, and victim's age are the most important while victim's gender, area of the crime, and victim's descent are the least important.

## Conclusion

**Sally Alakeel**

Our Logistic regression model would not be the best model to use in predicting whether a crime is violent or not violent. We had to limit the random samples to be able to provide better efficiency for this model. Since this model is more efficient for smaller datasets, it results in the predictors showing very weak significance since 14 out of our 22 predictors were found significant. This model also exhibits an extremely high error rate at 44.7% which means it results in incorrect predictions about half the time.

The random forest model provides more of an advantage since it works better with larger and more complex data. Our random forest model shows that it performed better than our logistic regression model. We found that the most important variables were, time of occurrence, victims age, and location. The test error rate was about 23-26% which is lower than the logistic regression model but also not ideal.

In conclusion, both models have a high error rate which could be a result of the dataset being flawed. Although neither model is ideal, the random forest model does perform better in predicting whether a crime is violent or not violent.

**References:**

**"Crime Data from 2010 to 2019"** *Kaggle,* October 11th, 2024, https://catalog.data.gov/dataset/crime-data-from-2010-to-2019 accessed 14 November 14, 2024