

➤ Baltimore City Crime Classification

Daniel E. de la Rosa, Moisés Díaz & Ean Jiménez

DataWizards

1 Introducción

Este ejercicio fue presentado como parte del Datathon 2022: Reto 3¹, un evento organizado por la Sociedad de Datos Dominicana (BIGDATA DO) con el objetivo de que los participantes desarrollen un modelo de machine learning con el que resolver el problema planteado.

A continuación se presenta el enfoque que se utilizó para resolver el problema. Partiendo en primer lugar de una definición y planteamiento del problema. Seguido de una exploración de los datos y la preparación de los mismos para el entrenamiento del modelo. Posteriormente, se presenta el modelo de machine learning y su evaluación. Finalmente, se presentan las conclusiones y recomendaciones de trabajos futuros, que podrían ser realizados con el objetivo de mejorar el modelo.

¹BIGDATA Dominicana (2022)

2 Definición del problema

El problema planteado consiste en predecir la categoría de un crimen en Baltimore City, Maryland, Estados Unidos, a partir de un par de características de los mismos. Para ello, se cuenta con un conjunto de datos¹ que contiene información sobre 236,370 crímenes ocurridos en Baltimore City, Maryland, Estados Unidos, entre el 1 de enero de 2017 y el 31 de diciembre de 2011. El conjunto de datos cuenta con 4 variables, las cuales se describen a continuación:

- **TARGET:** Categoría del crimen. Es la variable objetivo que se desea predecir. Es una variable categórica con 4 niveles: **ASALTO**, **HURTO**, **FALSO** y **ROBO**.
- **X:** Coordenada X del crimen. Es una variable numérica que representa la Longitud en grados decimales de la ubicación del crimen.
- **Y:** Coordenada Y del crimen. Es una variable numérica que representa la Latitud en grados decimales de la ubicación del crimen.
- **DATETIME:** Fecha y hora del crimen. Es una variable de tipo fecha y hora que representa la fecha, hora, minuto y segundo en que ocurrió el crimen.

¹ «Conjunto de datos» (2022)

3 Análisis preliminar de los datos

Tal como se mencionara con anterioridad, el objetivo del ejercicio es poder clasificar cada uno de los crímenes en una de las 4 categorías disponibles: i) ASALTO, ii) HURTO, iii) FALSO y iv) ROBO. LLegados a este punto, es importante mencionar tres aspectos importantes:

1. En Estados Unidos existe toda una metodología para clasificar los crímenes¹, que va más allá de las 4 categorías que se tienen en el conjunto de datos. Por lo que es posible que algunas categorías se encuentren combinadas en una sola categoría, que además fue traducida al español.
2. La categoría FALSO fue agregada al conjunto de datos para fines del concurso. Pero, en la realidad, no existe una categoría de crímenes que se llame FALSO en la metodología oficial.
3. Si bien las bases de concurso no especifican si los datos corresponden a Baltimore City o a Baltimore County, al realizar un análisis espacial de los datos, se observa que la mayoría de los registros, con excepción de los que corresponden a la categoría FALSO, se encuentran dentro de Baltimore City. Por tal razón, en lo anterior y lo adelante se hace referencia solo a Baltimore City.

3.1 Análisis exploratorio de datos

En la Tabla 3.1 que se incluye más adelante se presenta la distribución de registros por categoría. En la misma se observa que la categoría HURTO es la que tiene mayor número de registros, con un 39.11%, seguida de las categoría ASALTO y ROBO. En contraste, la categoría FALSO es la que tiene menor número de registros, con apenas el 9.82% del total. Sin embargo, estas dristribuciones pudieran ser ligeramente distintas si se analizan por año.

Tabla 3.1: Número y proporción de registros por categoría

Categoría	n	prop
HURTO	92,440	39.11
ASALTO	69,438	29.38
ROBO	51,282	21.70
FALSO	23,210	9.82

En términos de tiempo, por otro lado, se observa en la Figura 3.1 que el número de registros se reduce conforme pasan los años. Es decir que mientras 2017 representan cerca de 23.3% del total de registros, 2021 representa tan solo el 17.1% de estos. Este comportamiento no es consistente entre todas las categorías. Sin embargo, en ningún caso se observa una variación que pueda tener efectos en el análisis realizado.

¹Véase «Classifications of Crimes - FindLaw» (2022)

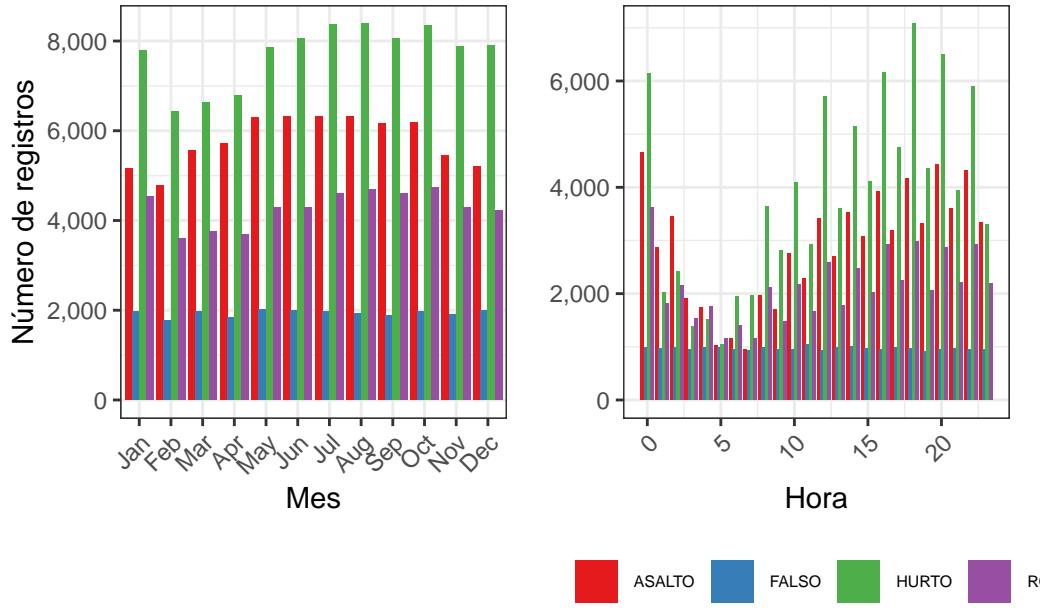


Figura 3.1: Número de registros por mes, hora y categoría

En la Figura 3.1 se presenta el número de registros por mes y categoría, así como el número de eventos por hora del día en que ocurren. En ambos casos se puede observar que los eventos muestran un comportamiento cíclico, con un pico en los meses junio-agosto y un mínimo en los meses diciembre-febrero. En el caso de las horas del día, se observa que los eventos ocurren principalmente entre las 16:00 y las 20:00 horas. Esto ocurre relativamente igual para todas las categorías, con la única excepción de la categoría FALSO, que muestra una distribución más uniforme a lo largo del día y del año.

Finalmente, en términos espaciales, en la Figura 3.2 se presenta un mapa de Baltimore City con los registros de delitos. En este mapa se puede observar que todas las categorías de delitos se distribuyen de manera relativamente uniforme a lo largo de la ciudad.

Todos los resultados anteriores se pueden resumir en el hecho de que si bien algunas de las variables revelan patrones en la data, ninguna de ellas proporciona elementos suficientes para realizar una separación apropiada de las clases.

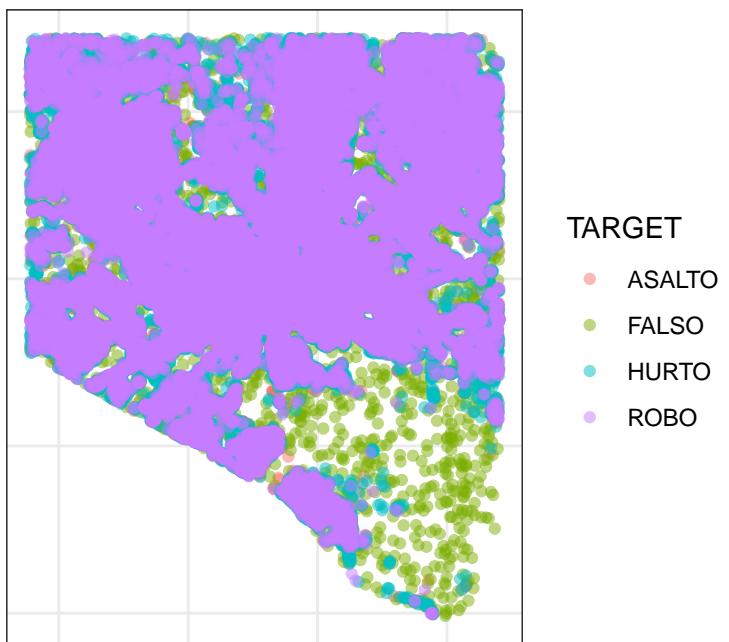


Figura 3.2: Mapa de Baltimore City con los registros de delitos

4 Modelado

4.1 Selección de variables

Dado el escenario descrito en la sección anterior, para la correcta clasificación de los crímenes se hace necesario encontrar y utilizar información adicional que permita segmentar los eventos. Es importante mencionar que como parte de las políticas del evento no está permitido utilizar datasets que contengan información adicional sobre específicamente estos eventos. Por lo tanto, se hace necesario encontrar información que pueda ser utilizada para la clasificación de los crímenes, pero que no sea específica de estos sucesos.

Para tales fines se procedió a consultar fuentes de información pública sobre Baltimore City. Un elemento importante lo constituye la tenencia de la información espacial en el dataset. Esta información permitió buscar datos geoespaciales que sirvieran al propósito del trabajo.

A continuación se presenta una lista no exhaustiva de las fuentes de información utilizadas.

4.1.1 «Open Baltimore» (NaN)

En este portal se puede encontrar información sobre la ciudad de Baltimore, incluyendo datos sobre la población, el empleo, la vivienda, la salud, la educación, el transporte, el medio ambiente, la seguridad, la economía, la cultura y el gobierno. Todas estas, variables que de acuerdo a «Variables affecting crime» (NaN) se consideran relevantes para la clasificación de los crímenes.

Toda esta información está disponible a nivel de ciudad o las llamadas *Community Statistical Areas* (CSA, 2020)¹. Estas áreas son agrupaciones de barrios que se utilizan para la medición de indicadores sociales. En Baltimore City existen 55 CSAs.

4.1.2 «Maryland's GIS Data Catalog» (NaN)

En este portal se encuentran, entre otras cosas, todos los archivos georreferenciados de dominio público referentes al estado de Maryland y sus demarcaciones.

4.1.3 «NIBRS – FBI» (NaN)

Esta página del FBI contiene información relacionada con crímenes, características de la víctima, tipo de ofensa, entre otras. Esta información se tiene a nivel agregado para todo el estado de Maryland, y no se tiene información geoespacial.

¹Véase «Neighborhood Health Profiles - Frequently Asked Questions | Baltimore City Health Department» (NaN)

4.2 Preprocesamiento de la data

4.2.1 Fecha

Como se mencionó anteriormente, la variable DATETIME contiene información sobre la fecha y hora en que ocurrió el evento. Para facilitar el análisis, se procedió a extraer la información de la fecha y hora en variables separadas. En el caso de la fecha, se extrajo el año, el mes, el día y el día de la semana. En el caso de la hora, se extrajo la hora y los minutos.

4.2.2 Análisis geoespacial

Para el análisis geoespacial se utilizó la información de las coordenadas geográficas de los eventos. Para ello fue necesario en primer lugar convertir los datos en un objeto espacial. Posteriormente, esto permitió realizar un análisis espacial y agragar otra información geoespacial a la data.

Especialmente se agregó la información relativa a las demarcaciones geopolíticas a las que pertenecen las coordenadas de los eventos. En este caso, se agregó la información de las CSAs de Baltimore City, así como el bloque² de la ciudad en el que se encuentra el evento.

Nótese que este ejercicio solo es factible para los datos que sean de tipo polígono. Para aquellos datos que no lo son, como es el caso de los puntos, por ejemplo paradas de autobus o estaciones de policía, se requiere de un análisis diferente. Por lo que se procedió a calcular la distancia mínima entre cada evento y cada punto de interés.

4.2.3 Data de los ATMS

Entendiendo que los delincuentes suelen ser maestros del oportunismo, nos pareció coherente analizar como la ubicación de los cajeros podía influir en los crímenes. Dicha data obtenida de mygeodata.cloud incluía las coordenadas de cada uno de los cajeros, por lo que utilizando el método de la distancia euclíadiana entre las coordenadas de cada crimen y las coordenadas de los cajeros, creamos la variable **Cercanía del cajero**, que al analizarla encontramos que es un 30% más probable que un crimen de tipo HURTO ocurre en las cercanías de un cajero

4.2.4 Data de llamadas del 911

Se procedio a obtener una base de datos del 2017 al 2021 sobre las llamadas que se hacían desde Baltimore al 911. En total unas 5 millones de llamadas. Para poder trabajar con esta data se creo una función que permitiera agrupar estos datos por zona, calculando así las siguientes variables de interés:

- Cantidad de llamadas por zona
- Cantidad de llamadas de alta prioridad por zona
- Porcentaje de llamadas de alta prioridad por zona
- Cantidad de llamadas de que no son emergencia por zona
- Porcentaje de llamadas de alta que no son emergencia por zona

²Según el Censo 2020.

Entre otros estadísticos como el mínimo, máximo, desviación estándar, etc. Ya con esta data agregada se procedió a incluir en nuestro set de datos

💡 Tip

Todo este ejercicio dió como resultado un dataset de 369 variables. Sin embargo, dado que estas son variables macro relacionadas con las variables de fecha y ubicación más que con el evento como tal, pocas de ellas aportan información relevante para la clasificación. De hecho, utilizano solo 100 de estas variables se puede llegar a resultados bastante similares.

De todos modos estas variables fueron sometidas a un análisis de importancia para determinar si eran relevantes para la clasificación. El resultado de este análisis se presenta en la sección de resultados.

4.3 Resultados

En el mundo del machine learning existen básicamente dos aproximaciones para el ejercicio de la clasificación multiclas:

1. La estrategia One vs. Rest (OvR) o One vs. All (OvA). En esta estrategia se entrena un modelo para cada clase, y se clasifica cada evento en la clase que tenga mayor probabilidad de pertenecer a ella.
2. La estrategia One vs. One (OvO). En esta estrategia se entrena un modelo para cada par de clases, y se clasifica cada evento en la clase que tenga mayor probabilidad de pertenecer a ella.

En un primer momento se pudiera pensar que para el ejercicio en cuestión la estrategia OvO es la más adecuada. Sin embargo, tras haber realizado el ejercicio, se llegó a la conclusión de que la estrategia OvR mostraba mejores resultados.

En tal sentido, se utilizó la estrategia OvR con un modelo XGBoost. El modelo se entrenó con un 77% de los datos y se validó con el 33% restante. El resultado de este ejercicio se presenta a continuación.

4.3.1 Importancia de las variables

Tabla 4.1: Importancia de las variables

Variable	Descripción	ROBO	HURTO	ASALTO	FALSO
viol20	Violent Crime Rate en esa zona	*			
priority-High- 1-sum-dummy- 911	Cantidad de llamadas al 911 de alta prioridad en esa zona		*		
Percent-of- Residents- Hispanic	Percent of Residents Hispanic en esa zona		*		

Variable	Descripción	ROBO	HURTO	ASALTO	FALSO
GeoIdBlockCensuzona		*	*	*	*
priority-High-1-mean-dummy-911	porcentaje de llamadas al 911 de alta prioridad en esa zona	*			
unempr10	Unemployment-Rate en esa zona		*		
pubtran18	Percent of Population that Uses Public Transportation to Get to Work en esa zona		*		
pwhite15	Percent of Residents White Caucasian(Non-Hispanic) en esa zona			*	
femhhs10	Percent of Female Headed Households with Children Under 18 en esa zona		*		*
hh40inc15	Percent of Households Earning 252C000				*
priority-Non-Emergency-0-sum-dummy-911	Cantidad de llamadas al 911 que no son emergencia en esa zona			*	*
hhm7518	Percent-of-Households-Earning-More-than-75%2C000				*

Esta estrategia permitió alcanzar un f1 score de 0.957 para los casos falsos, y tan solo 0.396 para los robos. Esto da como resultado un f1 score promedio de 0.0.548 para todas las categorías.

Es importante resaltar que estos scores se consiguen tras realizar un análisis de la curva ROC y determinar el umbral de clasificación que maximiza el F1 Score para cada uno de los modelos.

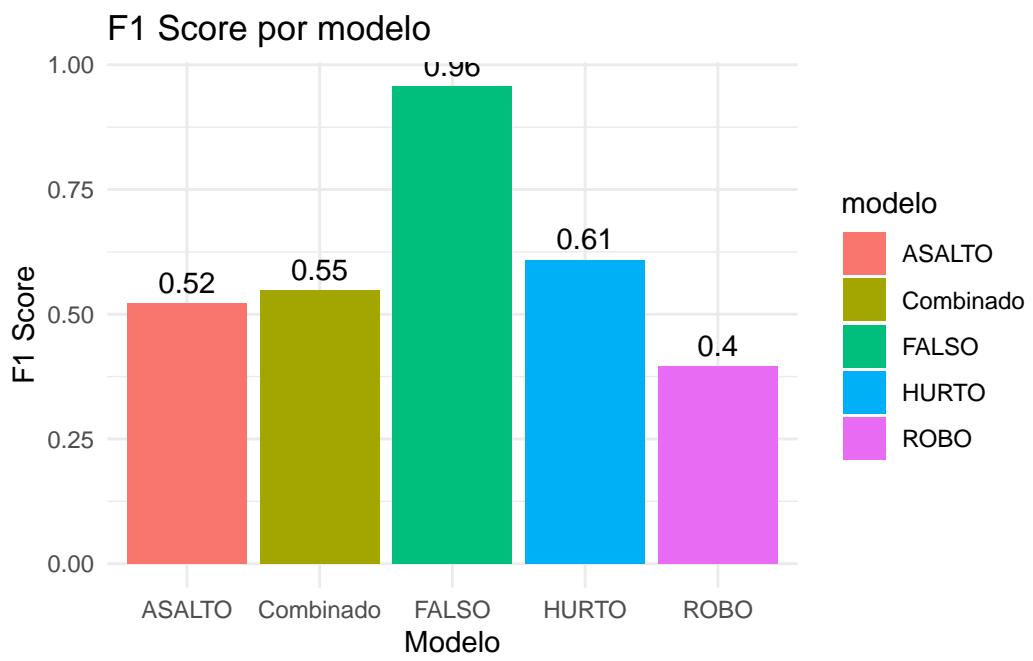


Figura 4.1: ?(caption)

5 Conclusiones

Tras la incorporación de las variables macro, se pudo observar que el modelo mejoró su performance. Sin embargo, no se pudo llegar a un modelo que clasifique con una precisión mayor al 60%. Esto se debe a que la data de los eventos es muy desbalanceada, y por lo tanto, el modelo tiende a clasificar todos los eventos como el tipo de crimen más frecuente.

También juega un papel el hecho de que no se pudiera agregar información adicional relacionada con el evento como tal. Por ejemplo, información sobre el sexo, edad, raza, etc. tanto de la víctima como del victimario.

De todos modos, la implementación de la estrategia OvR permitió obtener un modelo que clasifica los eventos con un Score F1 de 0.548.

Bibliografía

- BIGDATA Dominicana. 2022. «Bases del concurso». 2022. <https://github.com/BigDataDO/Datathon2022>.
- «Classifications of Crimes - FindLaw». 2022. 2022. <https://www.findlaw.com/criminal/criminal-law-basics/classifications-of-crimes.html>.
- «Conjunto de datos». 2022. 2022. <https://github.com/BigDataDO/Datathon2022>.
- «Maryland's GIS Data Catalog». NaN. NaN. <https://data imap.maryland.gov/>.
- «Neighborhood Health Profiles - Frequently Asked Questions | Baltimore City Health Department». NaN. NaN. <https://health.baltimorecity.gov/node/231>.
- «NIBRS – FBI». NaN. NaN. <https://www.fbi.gov/how-we-can-help-you/need-an-fbi-service-or-more-information/ucr/nibrs>.
- «Open Baltimore». NaN. NaN. <https://data.baltimorecity.gov/search>.
- «Variables affecting crime». NaN. NaN. <https://ucr.fbi.gov/nibrs/2012/resources/variables-affecting-crime>.