Análisis predictivo de violaciones por exceso de velocidad en zonas escolares de Chicago

Reporte Final del Proyecto

Profesor: Juan Castillo, PhD

Fecha: 23 de julio de 2025

Nombre del estudiante: Bryan León

Resumen

Este informe integra la propuesta inicial y los hallazgos clave del proyecto final de Modelos Predictivos. El objetivo fue construir un modelo predictivo que estime el volumen de violaciones por exceso de velocidad en zonas escolares de Chicago, utilizando variables temporales, geográficas y de identificación de cámaras. Se desarrollaron varios modelos demostrando una alta capacidad explicativa. Se identificaron variables clave como el año, la latitud y ciertas cámaras específicas.

1. Introducción y justificación

Las zonas escolares son entornos críticos para la seguridad vial, ya que involucran el tránsito de poblaciones vulnerables como niños y adolescentes. El cumplimiento de los límites de velocidad es esencial en estas áreas, y la tecnología de detección automática mediante cámaras ha permitido una mejor vigilancia. Sin embargo, los datos generados por estos sistemas son a menudo subutilizados. Este proyecto plantea que el análisis predictivo de dichos datos puede fortalecer las estrategias de prevención y control del tránsito, especialmente en el contexto urbano de una ciudad compleja como Chicago.

2. Objetivos

- Realizar un modelo predictivo que estime el número diario de violaciones por exceso de velocidad en zonas escolares.

- Identificar los factores temporales, geográficos y por cámara que más influyen en dichas violaciones.

- Evaluar el rendimiento de modelos de regresión lineal, Random Forest, Holt-Winter, Prophet y XGBoost

- Determinar la importancia relativa de cada variable predictora.

- Proporcionar recomendaciones basadas en evidencia para la gestión del tráfico escolar.

3. Marco teórico y antecedentes

La seguridad vial en zonas escolares representa una prioridad para los gobiernos locales, ya que los niños son considerados usuarios vulnerables de la vía. Con el objetivo de disminuir el riesgo de accidentes, muchas ciudades han optado por la instalación de cámaras automatizadas que registran violaciones por exceso de velocidad en horarios específicos. Estos registros permiten generar grandes volúmenes de datos que, adecuadamente tratados, pueden ser aprovechados para el desarrollo de modelos predictivos orientados a la prevención.

Desde una perspectiva analítica, las violaciones captadas por cámaras a lo largo del tiempo conforman una serie temporal, la cual puede presentar patrones de tendencia (incrementos o disminuciones sostenidas), estacionalidad (repetición cíclica de comportamientos semanales o anuales) y ruido aleatorio. Comprender y modelar estos patrones permite anticipar comportamientos futuros y optimizar la toma de decisiones en materia de fiscalización y seguridad vial.|

4. Dataset y fuentes

Se utilizó el dataset "Children's Safety Zone Program - Speed Camera Violations" del portal de datos abiertos de la Ciudad de Chicago. Incluye datos desde 2014, con información sobre fecha, código de cámara, latitud, longitud y número de violaciones por día. (Chicago, 2025)

Tipos de datos

Columna	Tipo de dato	Descripción
ADDRESS	Texto	Dirección específica donde está ubicada la cámara.
CAMERA ID	Texto	Identificador único de la cámara que captó la violación.
VIOLATION DATE	Fecha (tras conversión)	Fecha en la que ocurrió la violación.
VIOLATIONS	Entero	Número de violaciones registradas ese día por esa cámara.
X COORDINATE	Decimal	Coordenada X (proyección planar de ubicación).
Y COORDINATE	Decimal	Coordenada Y (proyección planar de ubicación).
LATITUDE	Decimal	Coordenada geográfica (latitud) del punto donde está ubicada la cámara.

LONGITUDE	Decimal	Coordenada geográfica (longitud).
LOCATION	Texto	Representación del punto geográfico (latitud, longitud) como texto.

5. Definición del Problema

En la ciudad de Chicago, las cámaras automatizadas instaladas en zonas escolares tienen la función de detectar y registrar violaciones por exceso de velocidad con el fin de proteger la integridad de los estudiantes y la comunidad. Sin embargo, el número de violaciones registradas presenta comportamientos variables a lo largo del tiempo y entre distintas ubicaciones. Estos patrones pueden estar influenciados por múltiples factores, como el día de la semana, el mes, la ubicación geográfica o cambios en la normativa.

La ausencia de un modelo predictivo limita la capacidad de las autoridades para anticipar aumentos en las violaciones, reasignar recursos de fiscalización, o evaluar el impacto de las cámaras instaladas. Por tanto, surge la necesidad de desarrollar un sistema que permita predecir con precisión el número de violaciones por exceso de velocidad en zonas escolares, usando datos históricos, variables temporales y espaciales.

El objetivo central de este proyecto es construir y comparar distintos modelos de predicción

6. Análisis Predictivo

En esta sección se describen los pasos realizados para el modelado predictivo.

6.a. Determinación de la base de datos

Se utilizó el conjunto de datos público **“Speed Camera Violations”** del portal de datos abiertos de la Ciudad de Chicago, el cual contiene registros diarios de violaciones por exceso de velocidad detectadas por cámaras instaladas en zonas escolares. La base incluye campos como la fecha de la violación, identificador de la cámara, ubicación geográfica (latitud y longitud) y la cantidad de infracciones por día. Esta fuente fue seleccionada por su amplitud temporal (desde 2014), nivel de detalle, frecuencia diaria y disponibilidad oficial actualizada. (Chicago, 2025)

6.b. Pre-procesamiento y limpieza

- Se convirtió la columna VIOLATION DATE al tipo de dato fecha.

- Se eliminaron registros con valores nulos en columnas esenciales como fecha, identificador de cámara, ubicación y número de violaciones.
- Se garantizó que los valores de VIOLATIONS fueran numéricos y se eliminaron los registros con valores negativos.
- Se eliminaron registros duplicados.

Además, se generaron variables temporales derivadas de la fecha: año, mes, día y día de la semana (WEEKDAY), las cuales resultaron útiles para detectar patrones cíclicos.

6.c. Análisis descriptivo

- Se elaboraron múltiples visualizaciones para comprender el comportamiento de las violaciones:
- Violaciones por año: Se observó un aumento significativo a partir de 2021, coincidente con cambios normativos que redujeron el umbral de velocidad sancionable. (Chicaco, 2021)
- Promedio por día de la semana: Se identificó mayor frecuencia de violaciones entre lunes y viernes, lo que concuerda con los días hábiles escolares.
- Top 10 cámaras: Se detectaron cámaras específicas con altos volúmenes de violaciones, lo que sugiere posibles zonas críticas de incumplimiento.
- Estas visualizaciones permitieron validar hipótesis y guiar la selección de variables para el modelado.

6.d. Selección de variables

Se definió como variable objetivo la columna VIOLATIONS, representando el número de infracciones diarias por cámara. Las variables predictoras seleccionadas fueron:

Temporales: YEAR, MONTH, DAY, WEEKDAY

Espaciales: LATITUDE, LONGITUDE

Categóricas: CAMERA ID, WEEKDAY (transformadas con codificación *OneHotEncoding*)

Además, para los modelos semanales y de series de tiempo, se crearon variables como número de semana, trimestre y agrupaciones semanales mediante resample.

6.e. Selección de modelos

Se implementaron y compararon diversos modelos representativos de tres enfoques distintos:

- Modelos estructurados:

Regresión Lineal: Usado como modelo base.

Random Forest: Con validación cruzada y optimización de hiperparámetros mediante GridSearchCV. Obtuvo un R^2 de 0.92 y MAPE de 45.71%.

XGBoost: Entrenado tanto con datos diarios como semanales, mejorando la robustez ante valores extremos.

- Modelos de series de tiempo:

Holt-Winters: Evaluado en tres configuraciones. El modelo semanal (estacionalidad de 52 semanas) fue el mejor dentro de esta familia (MAPE: 0.53%).

Prophet: Se aplicaron dos versiones: diaria (desde 2014) y semanal (desde 2021). Esta última alcanzó el mejor desempeño general con un MAPE de 0.34% y SMAPE de 27.89%, logrando capturar adecuadamente la estacionalidad diaria y semanal.

Modelo	MAE	RMSE	R^2	MAPE (%)	SMAPE (%)
Random Forest (Optimizado)	8.48	15.30	0.9246	45.71	30.56
XGBoost (Semanal)	7.89	13.90	0.9312	42.60	28.90
Holt-Winters (diario)	9.50	17.10	-	0.83	41.92
Holt-Winters (Mensual)	10.80	20.45	-	1.72	60.45
Holt-Winters (Semanal)	7.92	13.22	-	0.53	37.76
Prophet (Diario)	9.20	16.85	-	0.92	41.30
Prophet (Semanal)	6.55	12.44	-	0.34	27.89

7. Conclusiones

El análisis predictivo permite anticipar comportamientos recurrentes.
A partir de los datos históricos de violaciones captadas por cámaras de velocidad se logró identificar un modelo que puede ayudar a prever las violaciones por exceso de velocidad. Información clave para diseñar políticas públicas más efectivas en zonas escolares.

Los modelos de series de tiempo mostraron una mejor capacidad de ajuste que los modelos estructurados tradicionales.
Si bien algoritmos como Random Forest presento buenos resultados en términos de R^2, su

desempeño en métricas de error relativo como MAPE fue superado por modelos como Prophet especialmente cuando se trabajó con frecuencia semanal.

La calidad del preprocesamiento y la selección de variables fueron determinantes en el rendimiento del modelo.
La generación de variables temporales como WEEKDAY y la codificación adecuada de variables categóricas contribuyeron significativamente a mejorar la capacidad predictiva de los modelos.

El modelo Prophet semanal resultó ser la mejor alternativa para este caso de uso.
Con un MAPE de 0.34%, el modelo Prophet aplicado a datos semanales desde 2021 demostró un excelente ajuste a los ciclos reales observados en la serie, siendo además flexible, interpretable y fácil de actualizar.

Este estudio demuestra el valor del análisis de datos para apoyar decisiones con impacto social. Aplicar técnicas de ciencia de datos a problemas de seguridad vial no solo tiene valor académico, sino que también puede contribuir a proteger vidas y optimizar recursos públicos mediante acciones más focalizadas y basadas en evidencia.

8. Recomendaciones y Futuros estudios

Explorar variables externas como temperatura, feriados o tránsito.
La incorporación de variables exógenas podría mejorar la capacidad predictiva de los modelos. Por ejemplo, incluir información sobre clima, feriados escolares o eventos especiales ayudaría a explicar anomalías en la serie.

Integrar visualizaciones interactivas.
Para una mejor comunicación con audiencias no técnicas, se recomienda la construcción de dashboards en herramientas como Power BI o Tableau que permitan explorar violaciones por zona, fecha y cámara.

Estudiar el impacto de la instalación de nuevas cámaras.
Un análisis de series temporales interrumpidas o modelos causales permitiría estimar si la instalación de nuevas cámaras en ciertas zonas ha producido una disminución real en la velocidad o solo un desplazamiento del problema.

9.

Bibliografía

Chicaco, C. o. (2021). *Chicago Gov.* Retrieved from 2021 City of Chicaco Automated Enforment Program: https://www.chicago.gov/content/dam/city/depts/cdot/Red%20Light%20Cameras/2023/2021_AE_Report_FINAL.pdf

Chicago, C. o. (2025, July 06). *Transportation - Speed Camera Violations.* Retrieved from Chicago Data Portal: https://data.cityofchicago.org/Transportation/Speed-Camera-Violations/hhkd-xvj4/about_data

10. Anexos

GitHub Enlace: Bryan1696/Speed_violations: Estimar el número diario de violaciones por exceso de velocidad en zonas escolares.