Discovering the Relationship Between Crime Number and Weather in the Houston Area

Houston Omdena Local Chapter
Final Presentation
10/28/2023

Omdena local chapters

Chapter lead: Xuan Qin
https://omdena.com/chapters/

Mission

- Promote real-world AI through running open-source projects
- Provide case study-based education
- Provide AI services to local AI enthusiasts and businesses around the world
- Offline event

Vision

- Collaborate
- Network
- Deliver

Outline

Introduction

Data collection and preparation

Exploratory data analysis

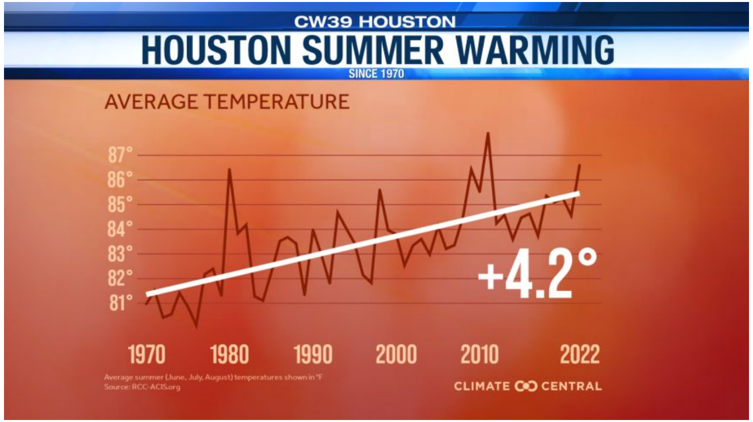
Model development and inference

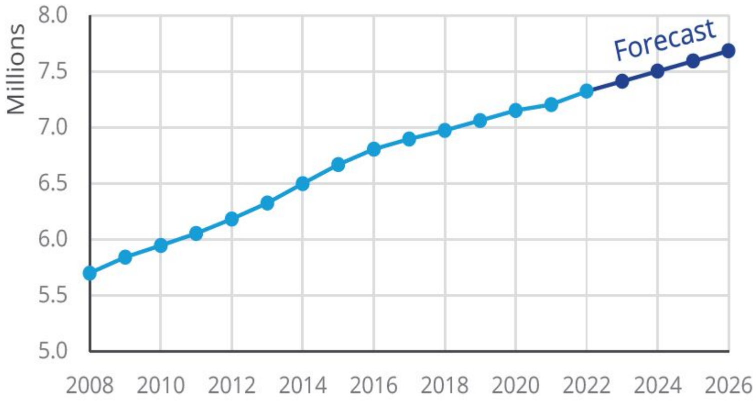
Model deployment

Concluding remarks

Background

- Rapid population growth posing social, safety, and economic challenges in Houston Metropolitan area.
- Global warming adds to the complexity, leading to more unpredictable weather events





Motivation

- To develop time-series analysis models that forecast crime numbers
- To investigate how much the predictive model can be improved by considering weather information
- To understand how weather factors influence the total crime number and specific types of crime numbers

Literature Review

- Prediction targets in crime prediction models
  - Location-based: Spatial relationships between crime and geographical factors
  - Type-based: Likelihood of theft, burglary, assault, etc. (Different types with different patterns/trends)
  - Temporal: Patterns and fluctuations in crime over different time periods
- Time-series analysis in crime prediction models
  - Forecasted the future values of daily, weekly, or monthly number of crime incidents within a specific time period based on historical data
  - Considered various temporal factors, such as Time, day, month, holidays, seasonality, trends, lagged variables, etc.
- Commonly correlated factors in crime prediction studies
  - Temperature, precipitation, humidity, wind speed, visibility, etc.

Data collection and processing

Goal: Collecting weather and crime data using available resources online and web-scraping. Clean, preprocess, and integrate data

Task co-leads: Miho, Ayesha

Collaborators:Tariq, Agata Kostrzewa, Andrew Yeh, Porselvi, Mary Aleta White, Márcia Cabral, Saleh Alhuraybi, Rimsha Sohail, Thanuja Stewart

Data collection

Site Name	Data Source URL	Cleaned data	Notes
City of Houston	https://www.houstontx.gov/police/cs/Monthly_Crime_Data_by_Street_and_Police_Beat.htm	https://dagshub.com/XuanQin/WeatherCrimeHouston/src/main/data/Cleaned/crime_jan2010_Jul2023.csv	Crime data (2010-01-01 to 2023-08-31)
Visual Crossing	https://www.visualcrossing.com/weather-history/Houston,TX/us	https://dagshub.com/XuanQin/WeatherCrimeHouston/src/main/data/Cleaned/cleaned_weather.csv	Weather data (2010-01-01 to 2023-08-31)
ArcGIS StoryMaps	https://cohgis-mycity.opendata.arcgis.com/datasets/MyCity::coh-police-beats/explore?location=29.840459%2C-95.387800%2C9.86	COH_POLICE_BEATS.zip	Houston GIS Data

Data processing

Dataset	**Data processing steps**
Crime Dataset	● 113 spreadsheet files (.xls and .xlsx) in 3 different formats were carefully examined and merged into 3 files. ● Removed leading and trailing spaces ● Consolidated values in multiple features into a single feature Ex. Offense Cont, Offenses, OffenseCount → Offense Count ● 3 files were merged into one ● Saved as csv
Visual Crossing	● 24 csv files were merged into one ● Saved as csv

Data Preprocessing Task

Task 2 was tasked with the following objective:

- ❑ Carry out an exploratory data analysis on crime and weather dataset
- ❑ Identify targets for the model to be developed
- ❑ Provide a machine readable data for the model development team

Task co-leads: Purva, Udo, Sabheen

Collaborators: Miho, Satish, Dihia, Ahmed, Mary, Minh, Tariq, Ayesha, Sahar, Marc

Task Approach

Exploratory Data Analysis:
- Univariate Analysis
- Bivariate Analysis
- Data viz

Data Cleaning:
- Missing values
- Quality Issues

Data Integration:
- Merge
- Harmonize

Transformation:
- Feature engineering

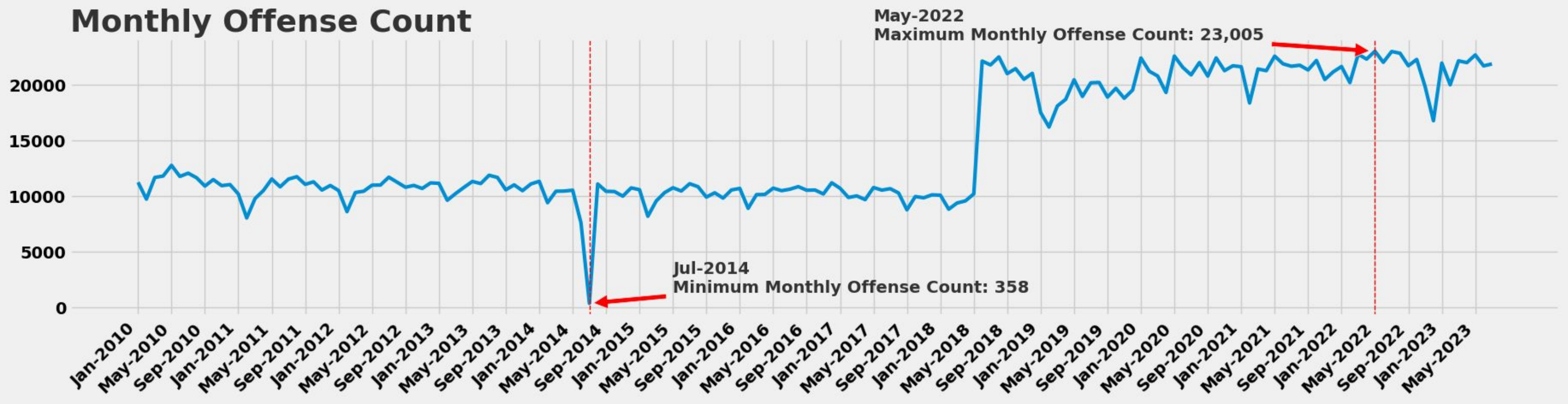
Documentation:
- All steps and versions
- Dagshub

Processed crime data
https://dagshub.com/XuanQin/WeatherCrimeHouston/src/main/data/Cleaned/all_crime_features_2010_2023_w_nibrs_class.csv

Univariate Analysis

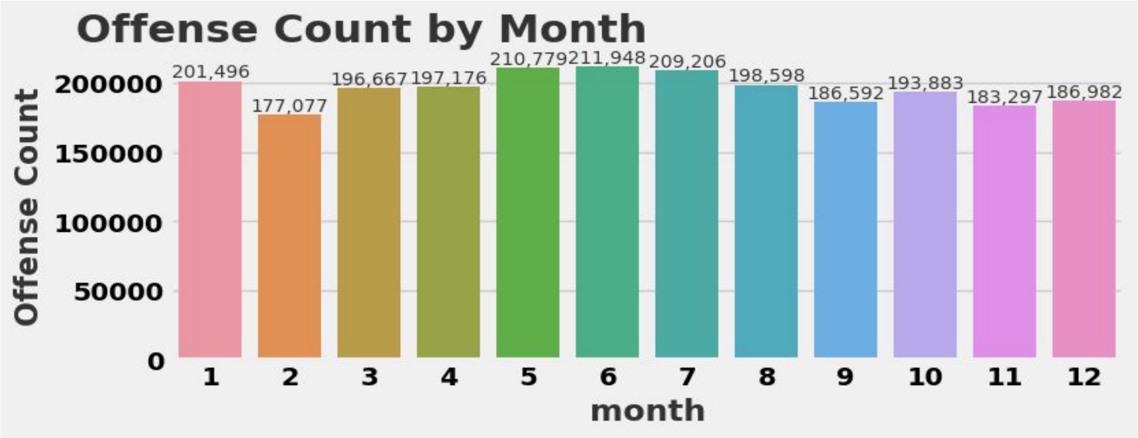
Weather data	Crime data
• No null value was detected. • 87% of the columns were numeric in data type. • 3 datetime columns exist. • The minimum humidity was found to be 18.1 and the maximum 99.6. • The minimum temperature in the dataset was 12.5 and the maximum was 110.5. • The minimum wind speed was recorded 0.9 and the maximum up to 40.9.	• Latitude and Longitude data available from 2022. • The raw data format changed in 2018-2019.

Monthly crime number history (2010-2023)



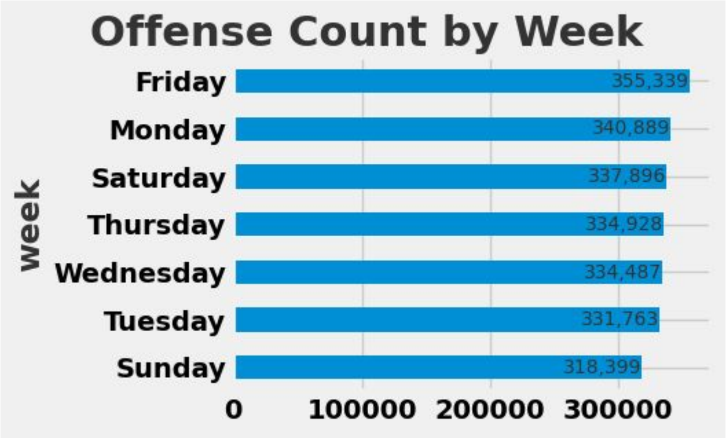
- Offense count increased after 2018. This could be due to the reporting change
- Offense Count increase after 2020 could have stemmed from something else, such as COVID-19

Crime count by month



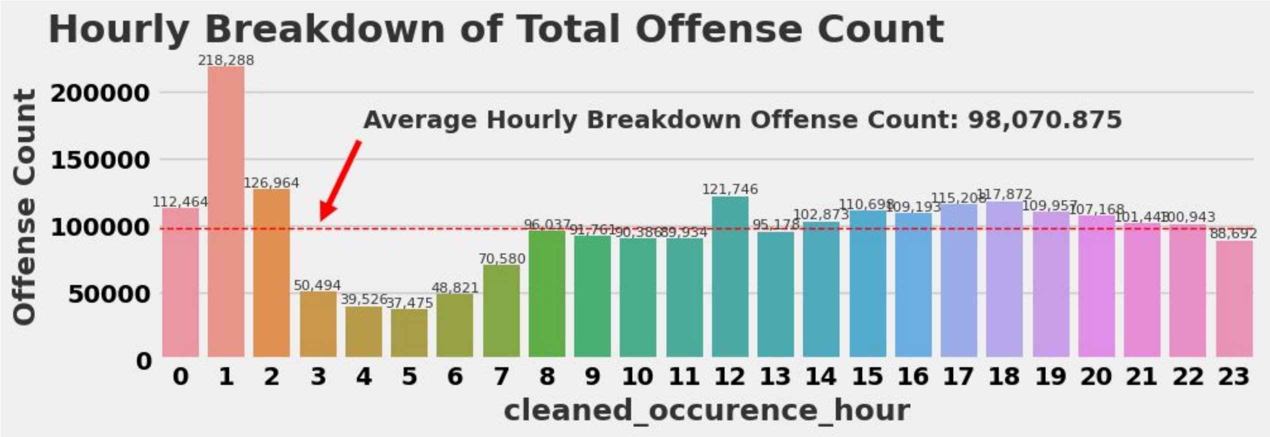
- Considering that February has only 28 days and that the data set covers January 2010 through July 2023, there appears to be no difference in the breakdown of the monthly totals.

Crime count by days of a week



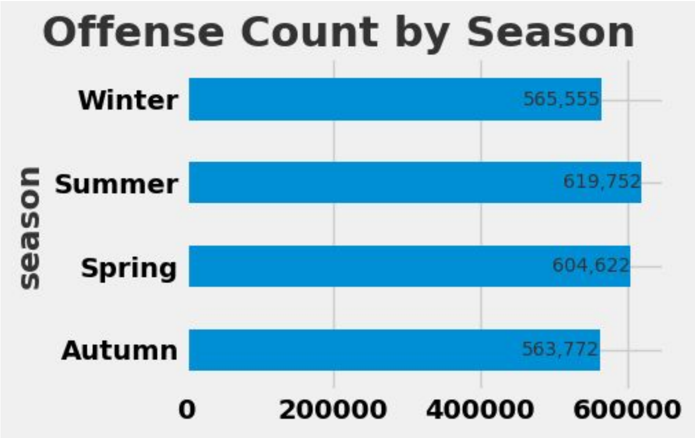
- Friday has the maximum 'Offense Count', while Sunday has the least (10% less than Friday)

Crime count by hour



- 1 am has the maximum total 'Offense Count', 223% of the average
- 5 am has the least number, 38% of the average

Crime count by season



December - February

June - August

March - May

September- November

- Summer has the maximum number 'Offense Count'
- Autumn has the minimum number 'Offense Count', 9% less than Summer

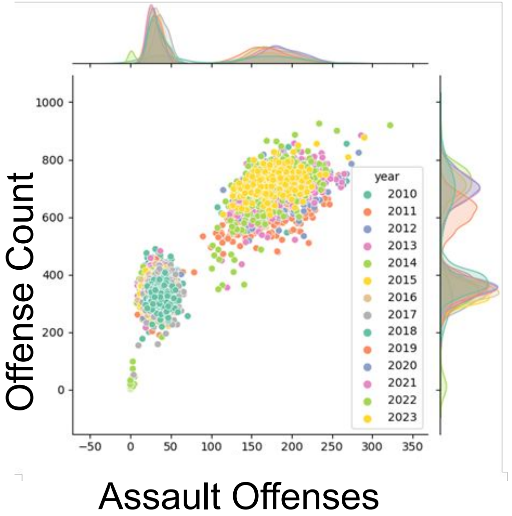
Label creation

Unify 36 specific crime types based on NIBRS code

Calculate total crime number and each crime type number per day

Merge with weather data based on 'date'

Daily crime count data for model development
https://dagshub.com/XuanQin/WeatherCrimeHouston/src/main/data/Cleaned/daily%20crime%20numbers%20and%20weather%20data%20for%20time%20series%20analysis.csv



Bivariate Analysis for daily crime count

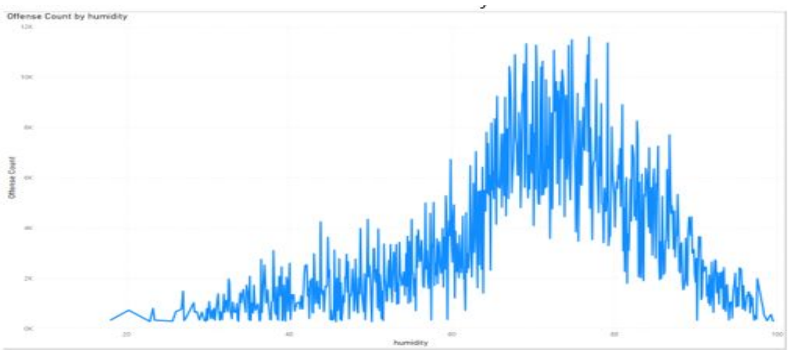


Fig. Humidity and Offense Count

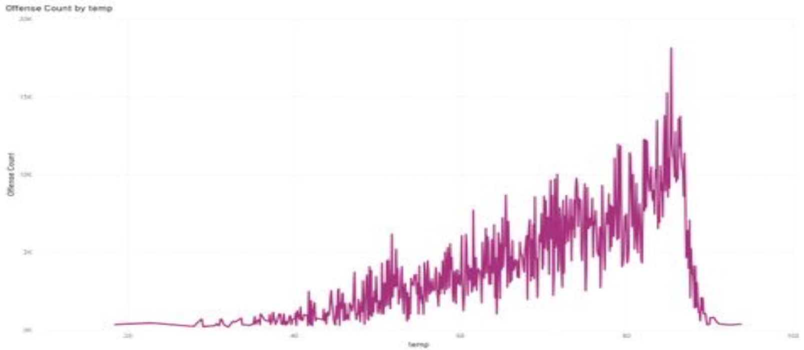


Fig. Temperature and Offense Count

- An important number of crimes is observed when the humidity is 60-80%.
- A significant number of crimes when the temp reaches 83-85 F.
- A very small crime rate is observed when it is snowy or rainy.

Bivariate Analysis (con’d)

Fig. Dew and Offense Count

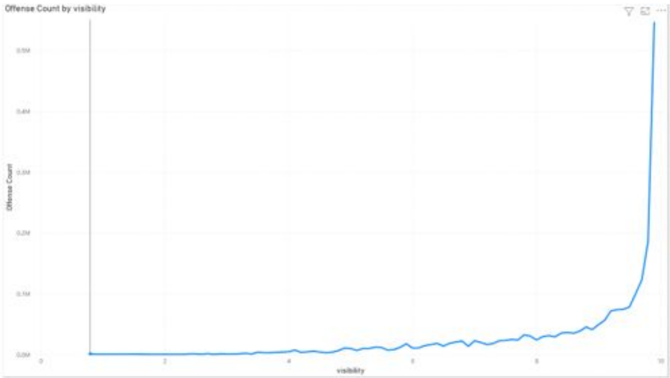


Fig. Visibility and Offense Count

- The crime rate is max when visibility is highest at 10. With reducing visibility, the crime rate is also reduced
- The dew value impacts the rate of crime similar to the temperature, suggesting positive correlation between dew and temperature

Bivariate analysis for temperature feature

Variable	F-Statistic	P-Value
Overcast	401.1740368	6.72E-86
Partiallycloudy	341.6052741	7.92E-74
Snow	186.4896464	1.04E-41
preciptype	65.22480953	2.31E-41
Clear	60.38188344	9.42E-15

Strongest Positive Correlation		Strongest Negative Correlation	
temp	**1.000000**	**temp**	**1.000000**
feelslike	0.993904	sealevelpressure	-0.594257
tempmin	0.977682	Overcast	-0.272813
tempmax	0.975805	cloudcover	-0.194305
feelslikemin	0.973140		
feelslikemax	0.972067		
dew	0.896572		

- The ANOVA tests indicate significant differences in temperature across various weather conditions.
- These categorical features can be included with temp in the model development

- Pearson's R suggest strong collinearity between feature temp and some temperature-related and dew features

Model development

***Objective*:** Provide data-backed insights for Houston PD to address weather-related crime spikes, enhancing public safety during extreme weather events.

***Goal*:** Develop a robust model to identify weather-related factors influencing crime rates, optimizing predictions within a 20% margin of error.

Task co-leads: Miho, Milan Kumar, Agata

Collaborators: Catalin, Tariq, Dihia, Porselvi, Satish Kumar

Time-series analysis model

Baseline models (Univariate):

- SARIMA
- LightGBM

Multivariate time series analysis:

- Random Forest
- XGBoost
- LSTM
- **VAR**
- **LightGBM**
- **EBM**

Target Variable: Daily Crime Count & daily count for specific crime types

Features:

- Weather
- Lagged Crime Count
- Temporal

Metrics:

- MAE
- MAPE
- RMSE
- R2

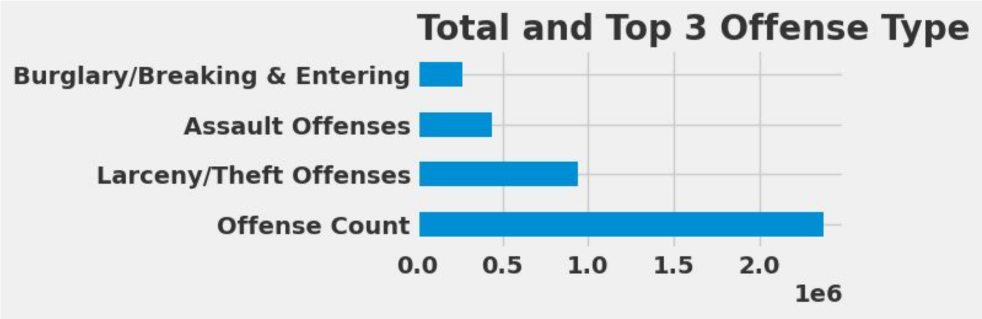
Vector Autoregressions (VAR)

Key advantages:

- **Multivariate Time Series Model**: VAR analyzes multiple time series variables simultaneously to capture complex relationships.
- **Lagged Variables Influence:** Each variable's current value depends on its own past values and the past values of other variables.
- **Bi-Directional Modeling:** VAR accounts for feedback loops, acknowledging that variables may influence each other in both directions.
- **Stationary Time Series Assumption:** Assumes that the statistical properties of the time series remain constant over time.
- **Future Value Prediction:** Utilizes historical data and lagged variables to make predictions about future values.

$$Y_{1,t} = \alpha_1 + \beta_{11,1} Y_{1,t-1} + \beta_{12,1} Y_{2,t-1} + \epsilon_{1,t}$$
$$Y_{2,t} = \alpha_2 + \beta_{21,1} Y_{1,t-1} + \beta_{22,1} Y_{2,t-1} + \epsilon_{2,t}$$

VAR Model: Feature Selection



- Total Offense Count and top 3 crime types were selected as target variables
- Weather-related features whose p-value is smaller than 0.05 in the Granger's Causality test were selected

Larceny/Theft Offenses & Weather related features

	p-value	causing	caused
47	0.0180	snow	Larceny/Theft Offenses
55	0.0546	windgust	Larceny/Theft Offenses
63	0.1003	winddir	Larceny/Theft Offenses
7	0.1342	tempmin	Larceny/Theft Offenses
11	0.1410	temp	Larceny/Theft Offenses

lagged_Assault Offenses & Weather related features

	p-value	causing	caused
25	0.0009	dew	lagged_Assault Offenses
85	0.0017	solarradiation	lagged_Assault Offenses
89	0.0017	solarenergy	lagged_Assault Offenses
5	0.0080	tempmin	lagged_Assault Offenses
1	0.0119	tempmax	lagged_Assault Offenses

lagged_Burglary/Breaking & Entering & Weather related features

	p-value	causing	caused
46	0.0353	snow	lagged_Burglary/Breaking & Entering
58	0.1452	windspeed	lagged_Burglary/Breaking & Entering
62	0.2739	winddir	lagged_Burglary/Breaking & Entering
54	0.3424	windgust	lagged_Burglary/Breaking & Entering
82	0.3966	moonphase	lagged_Burglary/Breaking & Entering

lagged_Offense Count & Weather related features

	p-value	causing	caused
76	0.0000	uvindex	lagged_Offense Count
84	0.0000	solarradiation	lagged_Offense Count
88	0.0000	solarenergy	lagged_Offense Count
28	0.0015	humidity	lagged_Offense Count
24	0.0018	dew	lagged_Offense Count

VAR Model: Feature Transformation & Inversion

- According to Augmented Dickey-Fuller (ADF) test result, Assault Offense and Offense Count were differentiated.

$$x_i'[j] = x_i[j] - x_i[j-1]$$

- For the model's accuracy, all features were min-max scaled.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- To compare forecast to actual values, inverse the forecast value.

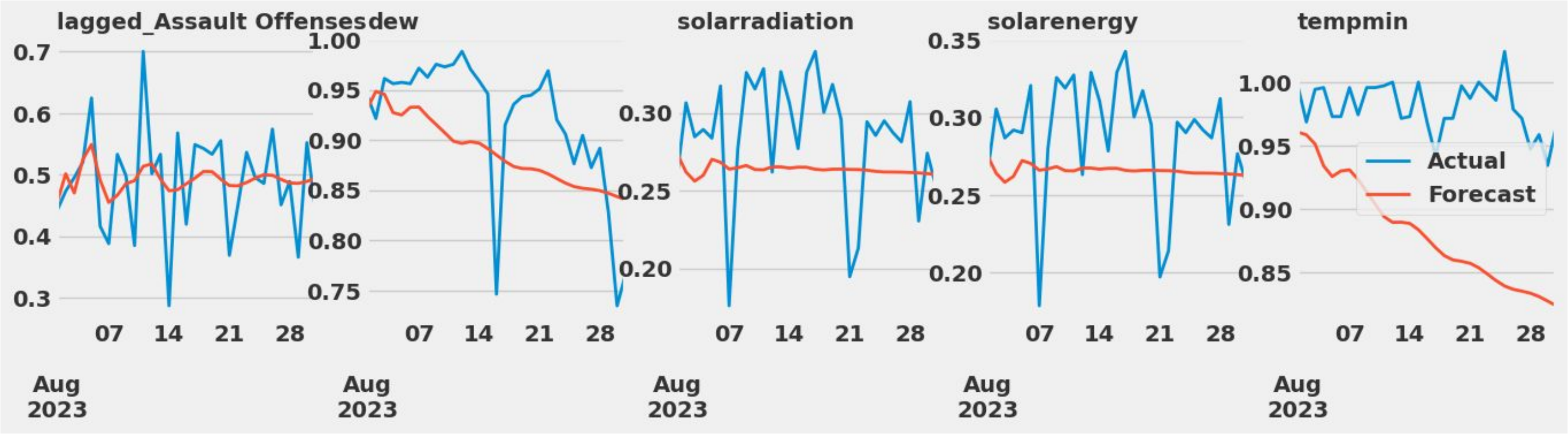
$$x = x' \cdot (\max(x) - \min(x)) + \min(x)$$

$$x_i[j] = \sum_{k=1}^{j} x_i'[k] + x_i[j-1]$$

VAR: Model evaluation

Training: 2010-01-01 to 2023-07-31
Forecast 2023 August

Model 1: Assault Offense, dew, solarradiation, solarenergy, tempmin



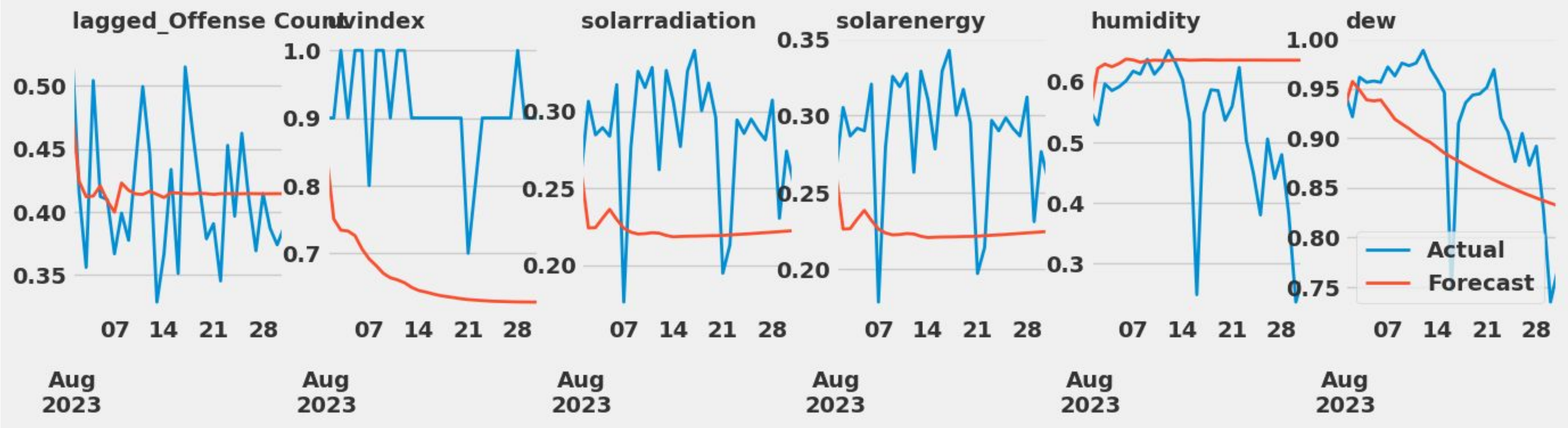
	MAE	MAPE	RMSE
7-days forecast	0.04	0.09	0.04
31-days forecast	0.06	0.11	0.13

- The Assault Offense forecast captures the trend
- Shown in the metrics table, as the forecasting horizon increases, the VAR model tend to lose its forecast accuracy

VAR: Model evaluation

Training: 2010-01-01 to 2023-07-31
Forecast 2023 August

Model 2: Offense Count, uvindex, solarradiation, solarenergy, humidity, dew



	MAE	MAPE	RMSE
7-days forecast	0.07	0.13	0.04
31-days forecast	0.1.	0.20	0.13

- Offense Count captured the actual trend, but it flattens after Aug 14.
- Shown in the metrics table, as the forecasting horizon increases, the VAR model tend to lose its forecast accuracy

LightGBM

Key advantages:

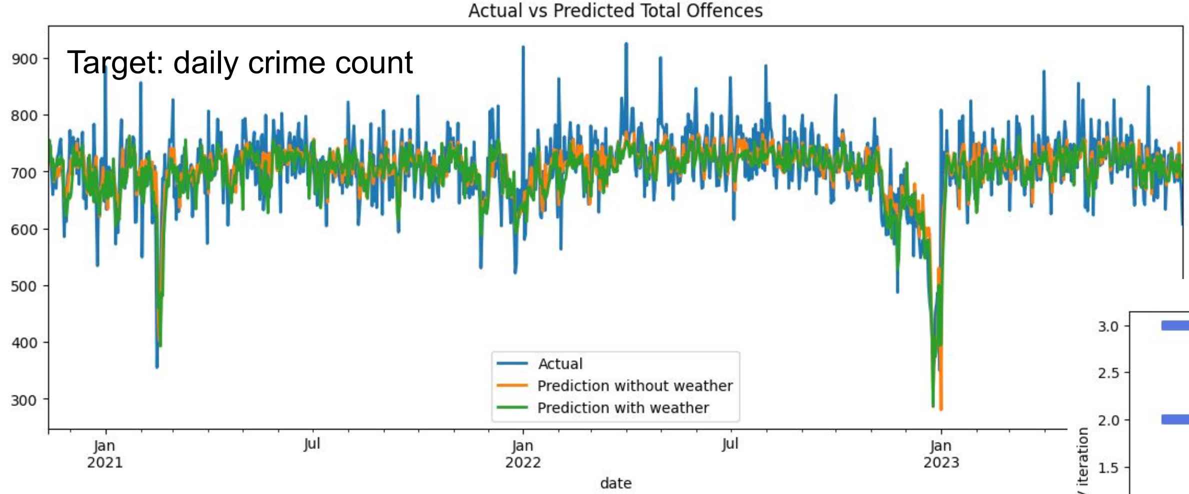
Boosting with Residuals: Enhances predictive power by focusing on the remaining errors from previous models

Effective Bagging: Utilizes bagging on both features and samples, ensuring robustness and stability.

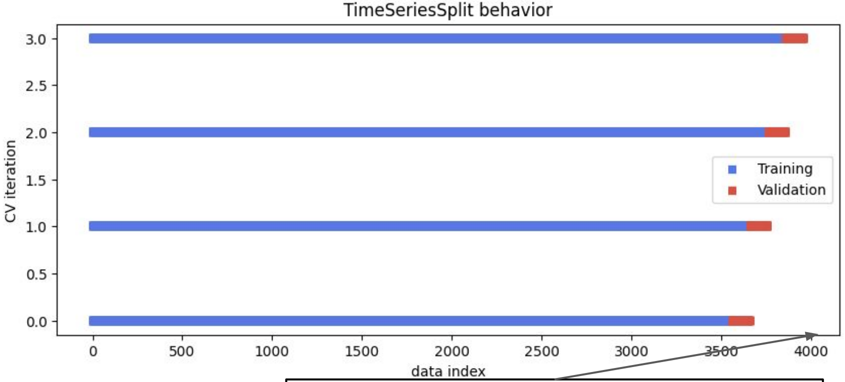
Enhanced Accuracy: Demonstrates superior accuracy in forecasting, especially with large datasets.

Efficiency: Boasts low memory usage and rapid training speed, ideal for real-time or resource-constrained environments.

LightGBM: Model evaluation



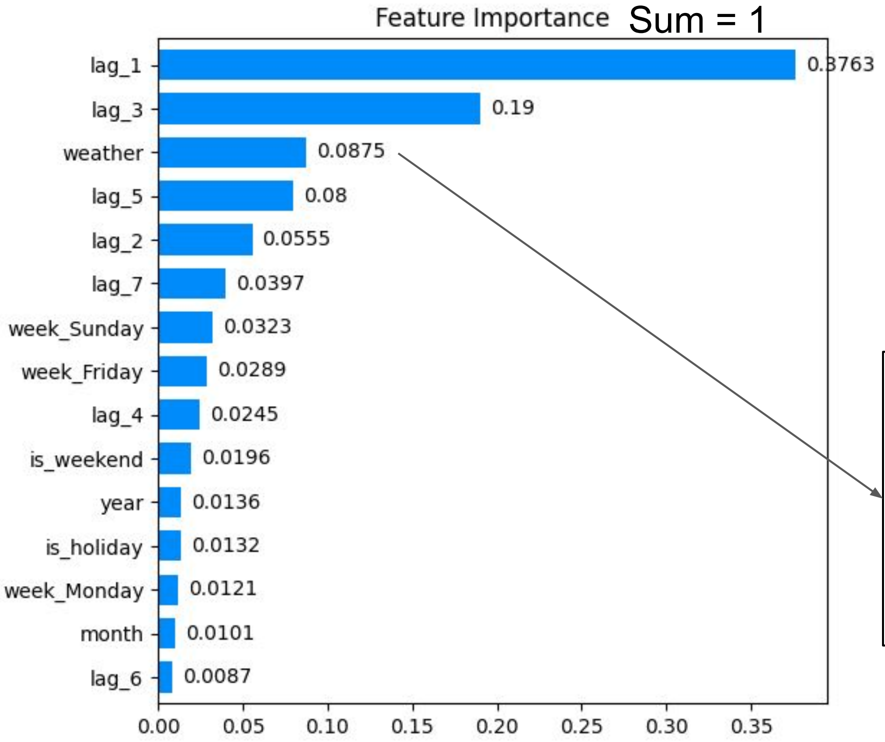
Train LightGBM models w/wo weather information using Grid search for hyperparameters and Time Series Split cross-validation.



Test metrics	MAE	MAPE	RMSE
w/o weather	38.40	0.056	52.96
w weather	37.86	0.054	51.53

Training with 80% of data

LightGBM: Feature importance

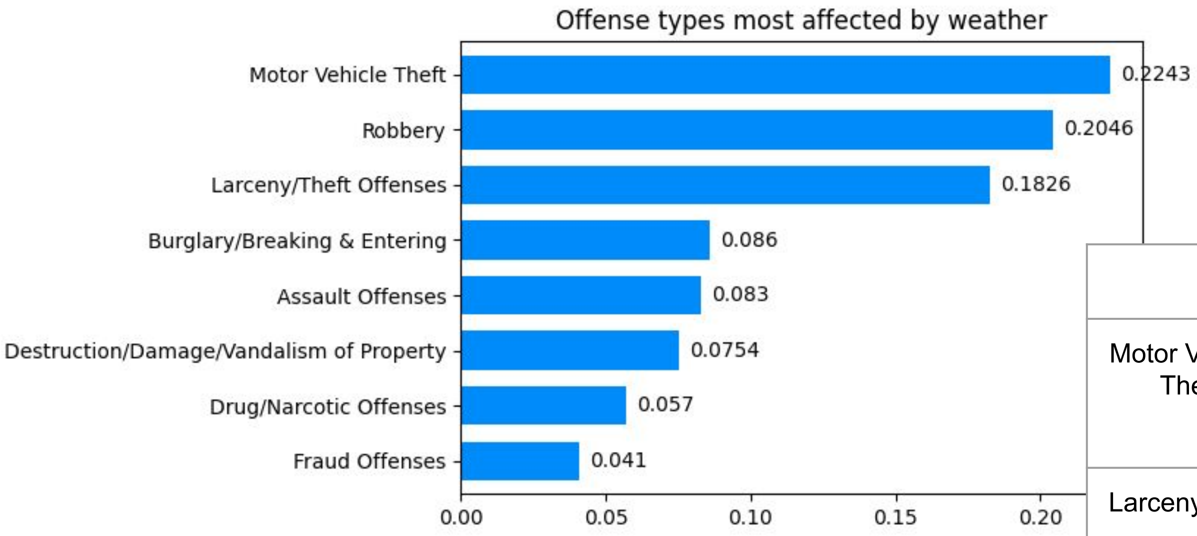


SHAP module is used to explain the model:

- Global and local interpretation
- Reliability and consistency
- Considering interactions
- Handling complex black-box models

- The highest feature importance in weather factors is tempmax (0.01)
- Sum up all feature importance of weather factors (temp, precip, conditions, moonphase, wind, solar, etc.)

Top crime types affected by weather



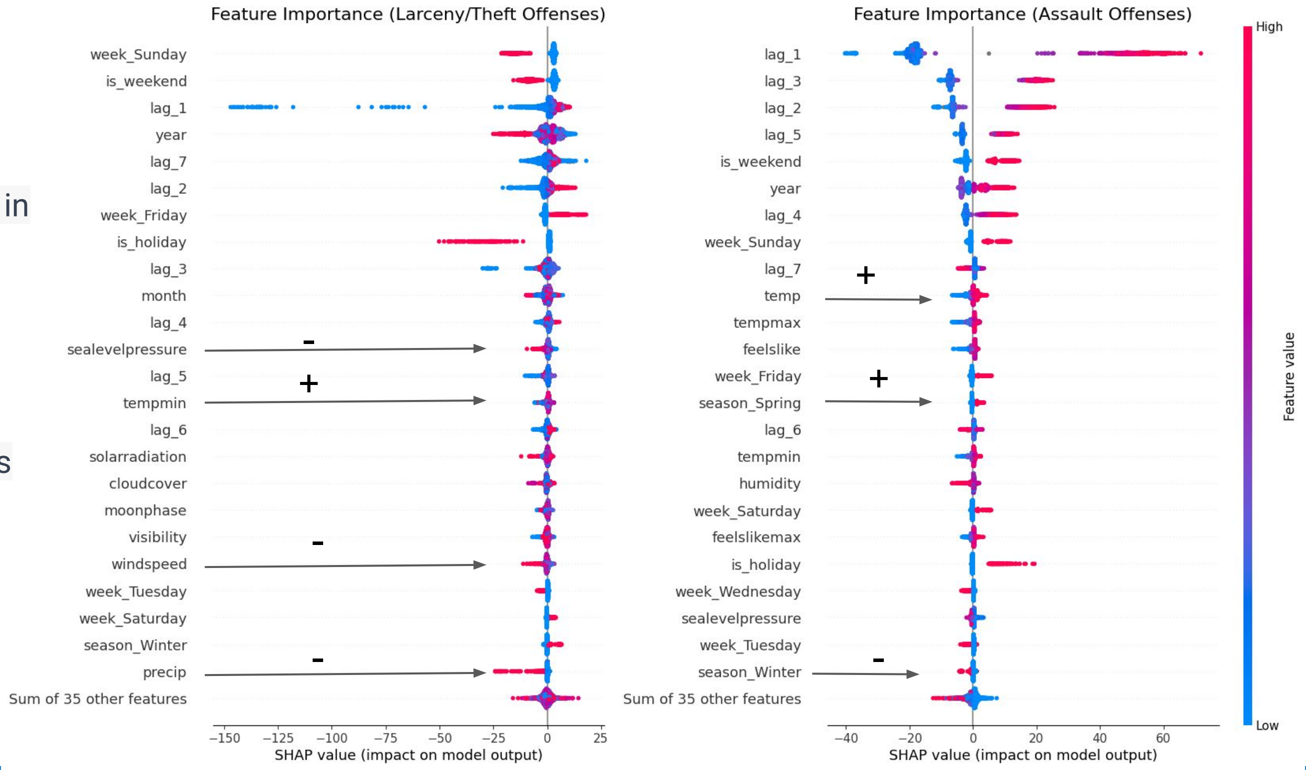
		MAE	MAPE	RMSE
Motor Vehicle Theft	w/o weather	9.44	0.188	11.96
	w weather	9.54	0.190	12.06
Larceny/Theft	w/o weather	19.25	0.108	25.31
	w weather	19.07	0.106	25.01
Assault Offenses	w/o weather	19.75	0.116	25.24
	w weather	18.55	0.108	24.17

Train LightGBM models w or w/o weather information using Grid search for hyperparameters and Time Series Split cross-validation for top 8 crime types.

Feature importance break down for crimes

- ***Precipitation and Wind speed***: Increasing levels of precipitation and wind speed correlate with a decrease in Larceny/Theft Offenses.

- ***Temperature-Related Factors***: Higher values of temperature-related features correspond to an increase in Assault Offenses.



Explainable boosting machine (EBM)

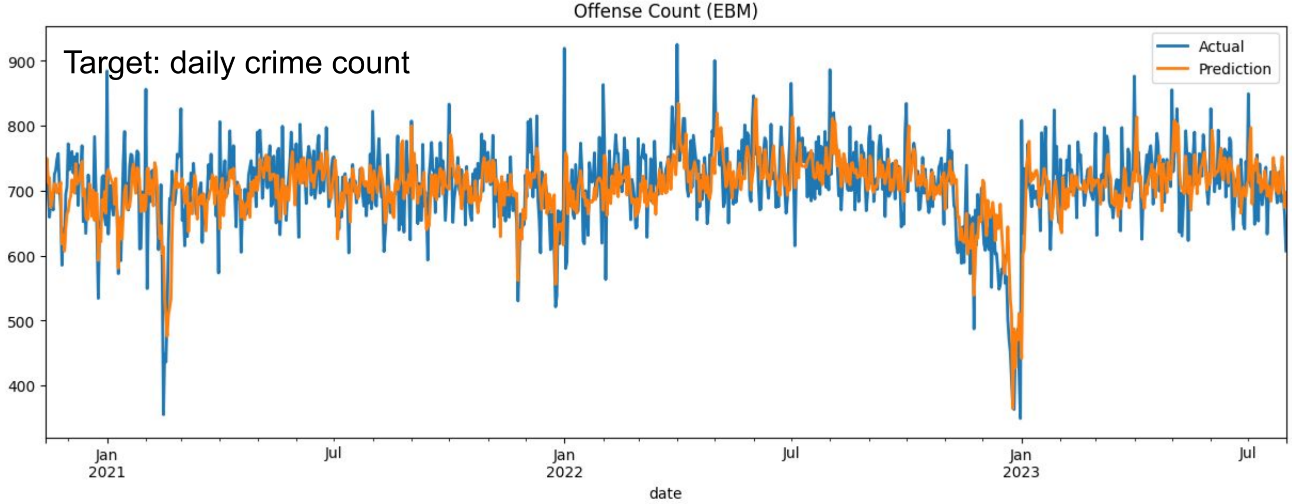
Key advantages:

- **Interpretable Machine Learning:** EBM provides clear, understandable insights, crucial for decision-making.
- **Sequential Feature Training:** It trains on one feature at a time, ensuring focus on individual variables.
- **Pairwise Interaction Consideration:** Considers interactions between features, capturing nuanced relationships.
- **Efficient Prediction:** Utilizes lookup tables for rapid predictions, making it suitable for real-time applications.
- **Visualize Feature Contributions:** EBM offers easy-to-understand visualizations for each feature impact.

$$g(E[y]) = \beta_0 + \sum f_i(x_i) + \sum f_{i,j}(x_i, x_j)$$

Generalized Additive Model

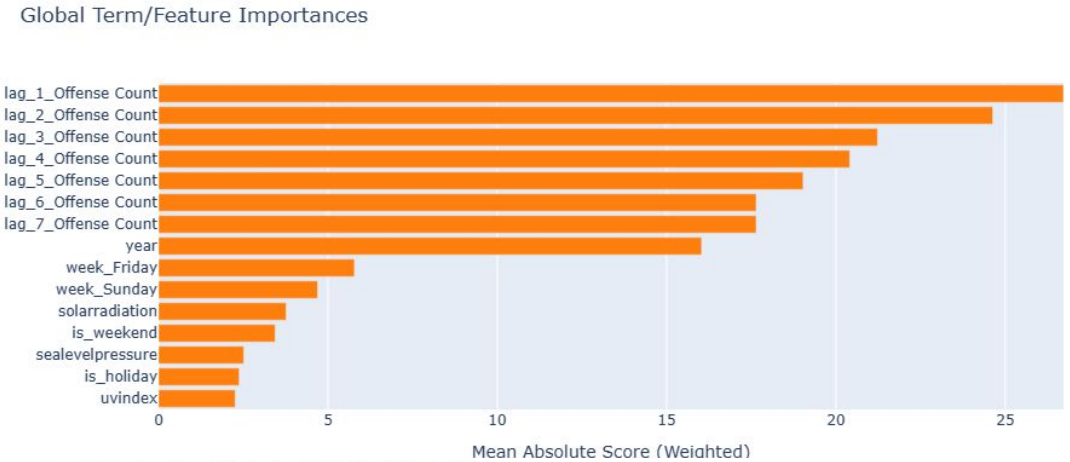
EBM: Model evaluation



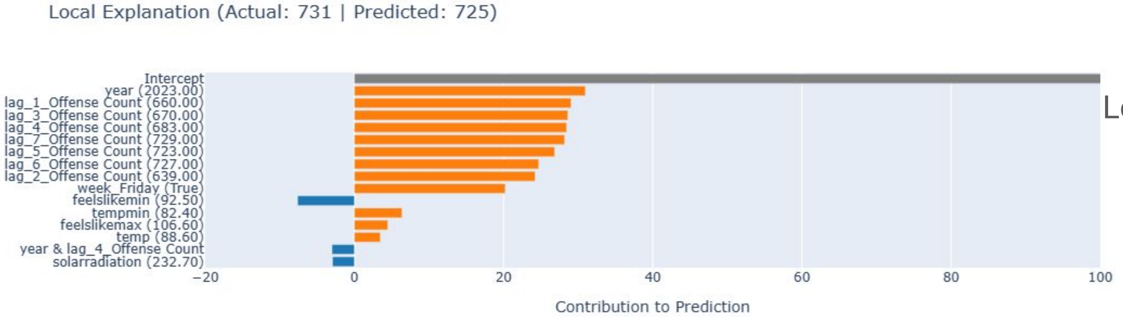
Cross-validate using Time Series Split on 80% of training data and Test on 20% of data

	MAE	MAPE	RMSE	R2
Daily crime count	38.18	0.055	51.44	0.36
Assault offenses	18.15	0.106	23.59	0.292

EBM: Feature importance and local explanation



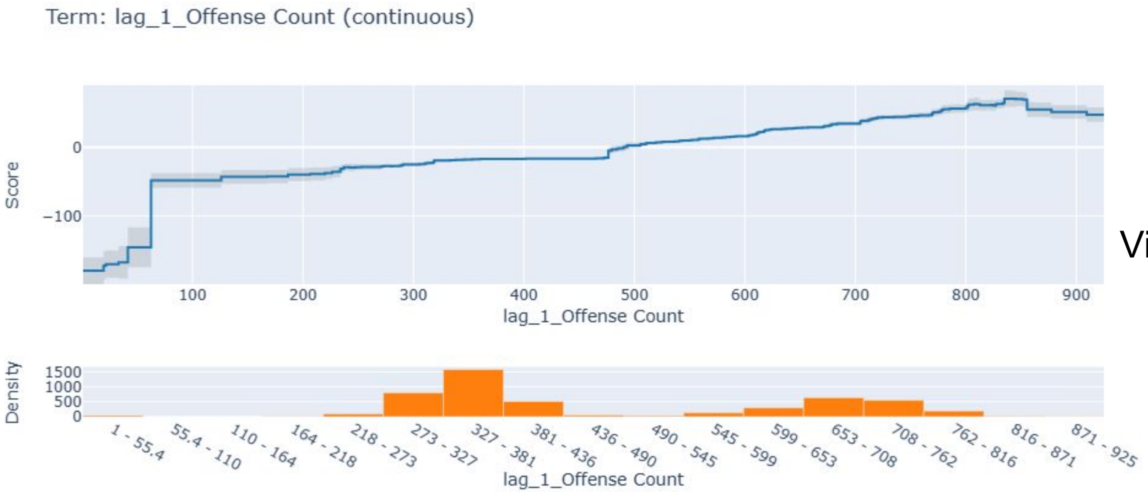
- *Target*: Daily Total Crime Count
- *Feature Importance*:
  - Lagged > Temporal > Weather
  - *Most Significant Weather Factors*:
    - Solar Radiation
    - Sea Level Pressure
    - UV Index



Local explanation for prediction on unseen data

- Intercept (475) dominates (towards mean)
- High importance does not ensure large prediction
- Feature interaction considered

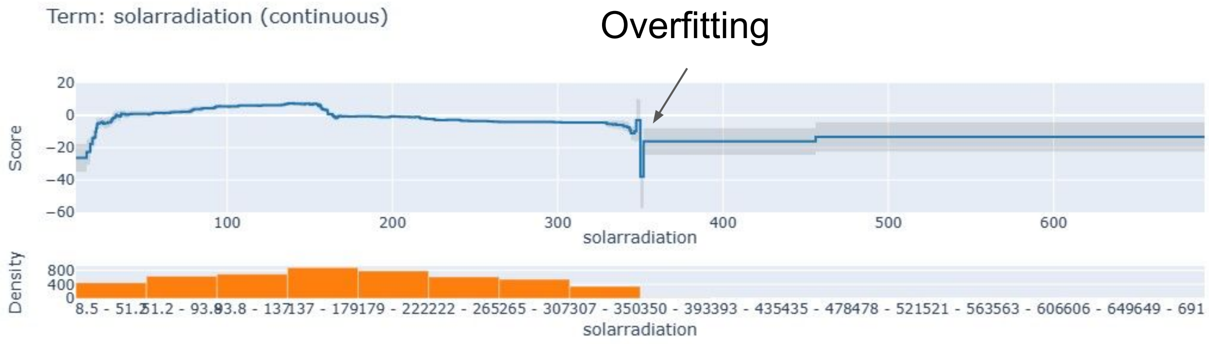
EBM: Historic and temporal factors

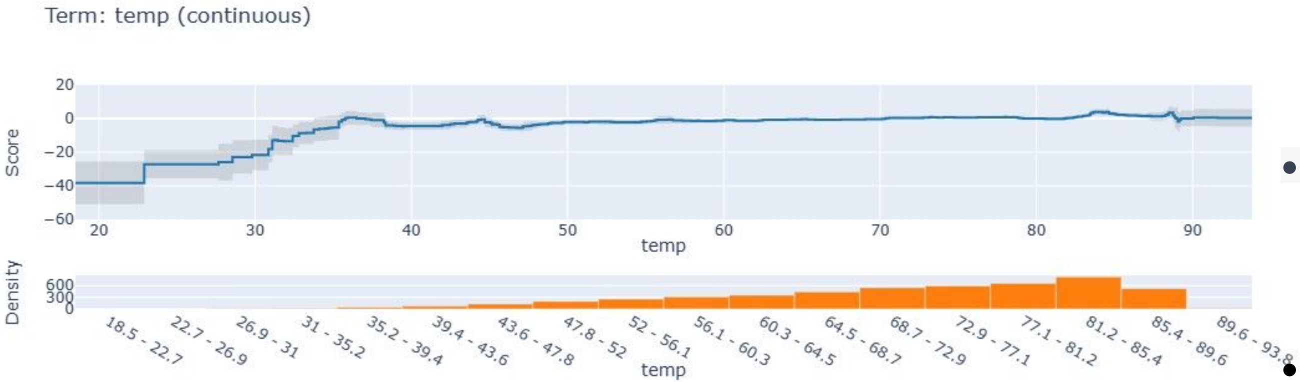


Visualize each feature as a lookup table

- Target with 1 to 7 lags can yield from -180 (±20) to 70 (±10)
- Yearly trend (gradual increase)
- Holiday has 44 fewer than non holiday
- Friday has 20 more
- Sunday has 16 fewer

EBM: Weather factors





- ***Solar Radiation***:
  - Increasing solar radiation initially leads to a rise in the daily crime count, followed by a subsequent decrease.
- ***Temperature***:
  - Higher temperatures result in a nonlinear increase in the daily crime count.
- ***Precipitation***:
  - Increasing precipitation leads to a decrease in the daily crime count, with the potential to reduce it by up to -100.
- ***Visibility***:
  - Reduced visibility, conversely, is associated with an increase in the daily crime count, potentially up to 20.

Outcome

Achievements

- Identified Crucial Weather-Driven Factors Influencing Crime Rates.
- Achieved Daily Crime Predictions with a 20% Margin of Error.

Model Precision

- Demonstrated Effective Forecasting Capabilities for Daily Crime Counts.

Feature Importance:

- Lagged Crime Count
- Temporal Features
- Weather Factors
  - Solar Radiation
  - Sea Level Pressure
  - UV Index
  - Temperature
  - Precipitation
  - Visibility

Model deployment

Goal: Deploy the machine learning model, built by our dedicated model development team, to provide an interactive platform for visualizing and forecasting crime rates in correlation with weather data.

Task co-leads: Chin Hao Zac, Milan Kumar, Saikrishna.

Collaborators: Sabheen

Gradio App

Key Advantages

- Fast, easy to set up
- Embedded in Jupyter Notebook
- Multi-model deployment

References: https://www.gradio.app/

Deployment process

- The web application features two tabs: 'Home Page' and 'EBM Prediction.'
- While our current deployment focuses on a single model, we plan to expand our model portfolio in the near future.
- Our model has been trained up to July 31, 2023, making data available for visualization in August 2023 and September 2023. Unfortunately, October data isn't accessible yet due to some missing values in the weather dataset.
- The visualization is currently organized by month, and in the near future, we will enhance it to provide segregation by both month and year.

Deployment process (continued)

```python
import gradio as gr
import matplotlib.pyplot as plt
import pandas as pd
import pickle
import numpy as np
import warnings
warnings.filterwarnings('ignore')
```

```python
def predict_august(data):
    # Predict using the loaded model
    with open("ebm.pkl", "rb") as model_file:
        ebm = pickle.load(model_file)
    predictions = ebm.predict(data)
    return predictions
```

```python
app2 = gr.Interface(
    fn=gradio_interface,
    inputs=gr.Dropdown(['August', 'September'], label="Select Month"),
    outputs=[gr.Plot(), gr.Dataframe()],
    live=True,
    title="Crime Rate Prediction and Visualization",
)
```

Deployment

Home Page | EBM Prediction

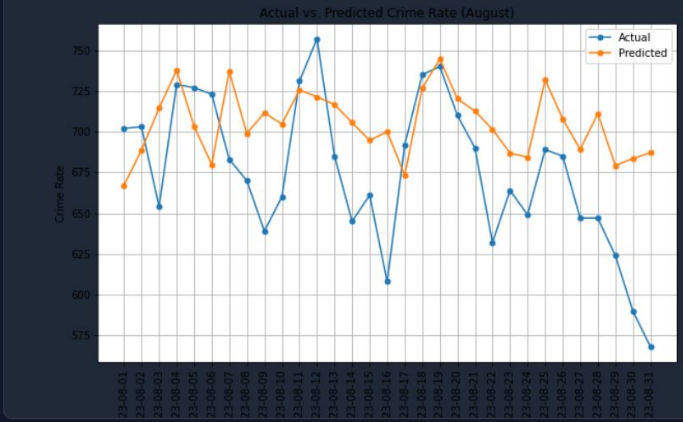
Crime Rate Prediction and Visualization

Select Month

August

Clear

output 0



output 1

Date	Predicted	Actual	Residual
2023-08-01	667	702	35
2023-08-02	689	703	14

Future work

- Data collection and processing: Creating data pipeline to automatically collect and process new data generated every month
- Exploratory data analysis: Feature selection and dimension reduction (PCA, SVD, and t-sne)
- Model development:
  - Higher granularity data and statistical testing preferred to investigate the relationship between climate change and crime rates.
  - Explore other data (employment) to improve crime forecast model
  - Explore the relationship between crime rate and location
  - Hypothesis testing on weather improves crime prediction
- Deployment: To write production-ready code that automatically retains model by schedule, using modular programming, CI/CD, and MLflow to track logs

Concluding remarks

- Crime count can be treated as a time-series prediction problem
- Weather factors slightly improve the forecast of daily crime count from the baseline model using only crime history by 3.6% in MAPE
- Glassbox models explain how factors of crime history, time, and weather factors yield the final prediction additively
- ML models demonstrates that certain crime types are affected by weather factors in a positive or negative way, and temperature affects crime rates to some degree

Repository link: https://dagshub.com/XuanQin/WeatherCrimeHouston

Email: qin.xuan1@gmail.com

http://www.linkedin.com/in/xuanqin

Thank you!

Omdena Houston