The Impact of Weather Conditions on Crime Rates in Chicago

***IST 718: BIG DATA ANALYTICS (Prof.* Michael Mulligan)**

PREPARED BY

SHOUMIK REDDY KUMBHAM

skumbham@syr.edu

BHANU KIRAN GARIKIPATI

bgariki@syr.edu

(GROUP 3)

I. Project Overview

Understanding the relationship between weather patterns and crime rates is crucial for effective urban planning and law enforcement strategies. In this project, we delve into the intricate dynamics between weather conditions and crime occurrences in Chicago.

Chicago, known for its diverse climate, serves as an ideal case study to explore how variations in weather parameters influence different types of crimes across various neighborhoods.

Our Main aims:

(a) **Analyze:**

1. **Crime rates in Chicago:** Conduct a comprehensive analysis to examine how different weather conditions correlate with overall crime rates in Chicago. Identify any significant patterns or trends indicating the influence of weather on the frequency and severity of criminal activities.
2. **Specific Crime types:** Delve deeper into the relationship between weather and specific types of crimes prevalent in Chicago. Analyze how weather conditions may affect the occurrence and characteristics of various crime categories, such as property crimes (e.g., theft, burglary) and violent offenses (e.g., assault, homicide).

(b) **Predict:**

1. **Crime Counts:** Build regression models to predict the overall volume of criminal incidents in Chicago based on prevailing weather conditions. This enables law enforcement agencies and policymakers to anticipate fluctuations in crime rates and allocate resources accordingly.
2. **Hotspots:** Employ spatial analysis techniques in conjunction with weather data to identify crime hotspots within the city. By integrating wards with machine learning algorithms, predictive models can pinpoint areas prone to heightened criminal activity under specific weather conditions, facilitating proactive intervention strategies.

II. Data preprocessing

a) We start by procuring (from crime data from Chicago open data portal and weather data from visual crossing) and loading the datasets into our analysis environment and inspecting their basic properties, such as the variables, and data types.

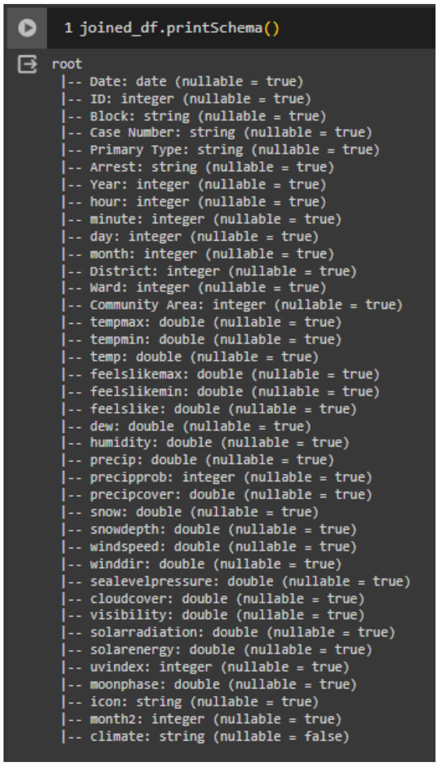
b) Ensure uniformity in date representations across datasets by standardizing date formats.

c) Identify and remove unnecessary columns such as Police Beat and Expired Ward Numbers from the datasets to streamline data processing.

d) Introduce new columns as needed to enhance data analysis and modeling, incorporating relevant information that may be missing from the original datasets.

e) Eliminate rows containing null values in critical fields such as Police Beats, Expired Ward Numbers, and Location Coordinates to maintain data integrity and accuracy.

f) Merge datasets based on the Date field to consolidate information from multiple sources and facilitate comprehensive analysis of the relationship between weather and crime rates over time.



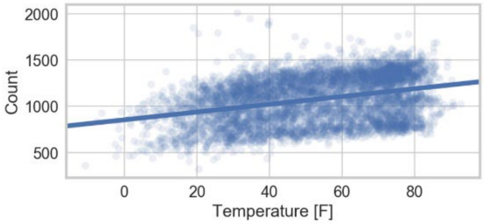
```
joined_df.printSchema()
```

```
root
 |-- Date: date (nullable = true)
 |-- ID: integer (nullable = true)
 |-- Block: string (nullable = true)
 |-- Case Number: string (nullable = true)
 |-- Primary Type: string (nullable = true)
 |-- Arrest: string (nullable = true)
 |-- Year: integer (nullable = true)
 |-- hour: integer (nullable = true)
 |-- minute: integer (nullable = true)
 |-- day: integer (nullable = true)
 |-- month: integer (nullable = true)
 |-- District: integer (nullable = true)
 |-- Ward: integer (nullable = true)
 |-- Community Area: integer (nullable = true)
 |-- tempmax: double (nullable = true)
 |-- tempmin: double (nullable = true)
 |-- temp: double (nullable = true)
 |-- feelslikemax: double (nullable = true)
 |-- feelslikemin: double (nullable = true)
 |-- feelslike: double (nullable = true)
 |-- dew: double (nullable = true)
 |-- humidity: double (nullable = true)
 |-- precip: double (nullable = true)
 |-- precipprob: integer (nullable = true)
 |-- precipcover: double (nullable = true)
 |-- snow: double (nullable = true)
 |-- snowdepth: double (nullable = true)
 |-- windspeed: double (nullable = true)
 |-- winddir: double (nullable = true)
 |-- sealevelpressure: double (nullable = true)
 |-- cloudcover: double (nullable = true)
 |-- visibility: double (nullable = true)
 |-- solarradiation: double (nullable = true)
 |-- solarenergy: double (nullable = true)
 |-- uvindex: integer (nullable = true)
 |-- moonphase: double (nullable = true)
 |-- icon: string (nullable = true)
 |-- month2: integer (nullable = true)
 |-- climate: string (nullable = false)
```

III. Data Analysis

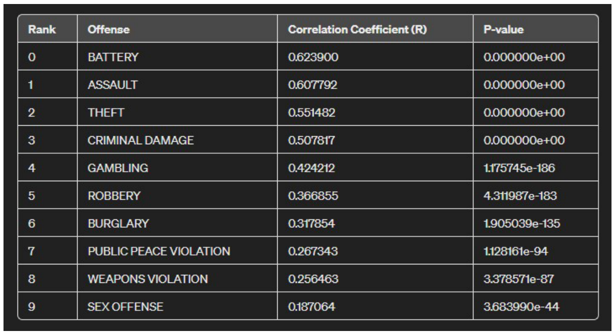
At this juncture, our primary objective is to uncover correlations between weather patterns and crime rates, crime types in Chicago, laying the groundwork for predictive modeling. By scrutinizing the preprocessed datasets, we aim to identify meaningful relationships and dependencies that can inform our prediction algorithms.

a) **The Impact of Weather on Crime Rates:**
Crime counts and maximum temperatures both exhibit seasonal patterns, with peaks generally aligning with warmer months and troughs in cooler months, suggesting a possible correlation where higher temperatures may be associated with increased crime counts.



b) **The Impact of Weather on Crime Rates:**

Conducted regression analyses to assess the impact of temperature on various crime types by fitting a linear regression model for each category using temperature as the predictor variable. Crimes such as "BATTERY", "ASSAULT", and "THEFT" have very significant p-values (much less than 0.05), indicating strong evidence against the null hypothesis (no relationship).



Rank	Offense	Correlation Coefficient (R)	P-value
0	BATTERY	0.623900	0.000000e+00
1	ASSAULT	0.607792	0.000000e+00
2	THEFT	0.551482	0.000000e+00
3	CRIMINAL DAMAGE	0.507817	0.000000e+00
4	GAMBLING	0.424212	1.175745e-186
5	ROBBERY	0.366855	4.311987e-183
6	BURGLARY	0.317854	1.905039e-135
7	PUBLIC PEACE VIOLATION	0.267343	1.128161e-94
8	WEAPONS VIOLATION	0.256463	3.378571e-87
9	SEX OFFENSE	0.187064	3.683990e-44

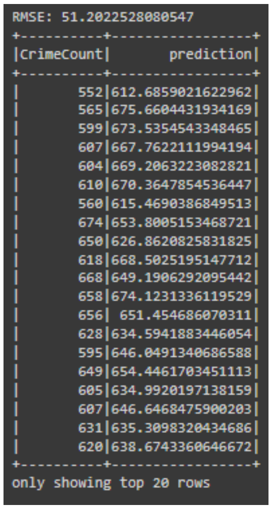
IV. Predictions

Building upon the insights gleaned from our data analysis regarding the impact of weather on crime rates in Chicago, we now transition to the pivotal phase of predictive modeling. Leveraging the observed correlations between weather variables and crime occurrences, our objective is to develop robust machine learning models capable of forecasting future crime counts, types, and hotspots based on prevailing weather conditions.

a) **Crime Counts:**

After assembling the features and preparing the data, we employed a Random Forest Regressor model to predict daily crime counts in Chicago. The model utilized various meteorological features such as maximum temperature, minimum temperature, precipitation, humidity, wind speed, and others to make predictions. We constructed a pipeline that integrated a Vector Assembler to concatenate the input features into a single vector and a Random Forest Regressor for training and prediction. The dataset was split into training and test sets with a ratio of 80:20, respectively, to assess model performance.

Following model training, we evaluated its performance using the Root Mean Squared Error (RMSE) metric. The resulting RMSE value was found to be 51.20, indicating the average deviation between the actual and predicted crime counts. This evaluation metric provides a measure of the model's accuracy in predicting crime counts based on weather variables.



```
RMSE: 51.2022528080547
+----------+-----------------+
|CrimeCount|       prediction|
+----------+-----------------+
|       552|612.6859021622962|
|       565|675.6604431934169|
|       599|673.5354543348465|
|       607|667.7622111994194|
|       604|669.2063223082821|
|       610|670.3647854536447|
|       560|615.4690386849513|
|       674|653.8005153468721|
|       650|626.8620825831825|
|       618|668.5025195147712|
|       668|649.1906292095442|
|       658|674.1231336119529|
|       656| 651.454686070311|
|       628|634.5941883446054|
|       595|646.0491340686588|
|       649|654.4461703451113|
|       605|634.9920197138159|
|       607|646.6468475900203|
|       631|635.3098320434686|
|       620|638.6743360646672|
+----------+-----------------+
only showing top 20 rows
```

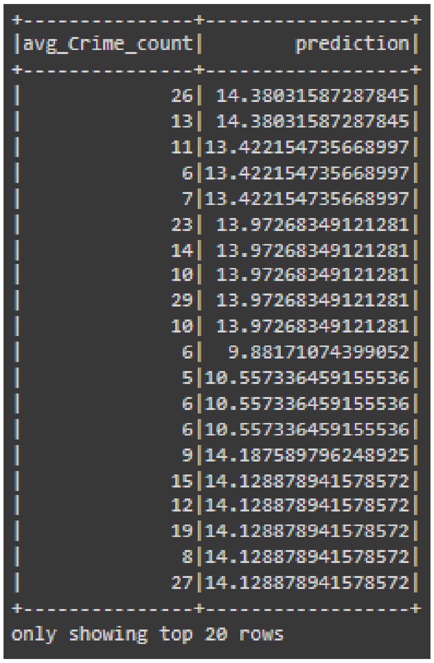
Upon examining the prediction results, we observed that the model's predictions closely align with the actual crime counts, albeit with some variance. The table displays a sample of predicted crime counts alongside their corresponding actual values.

b) **Hotspots:**

For the hotspot prediction using weather data, we adopt a ward-based approach, considering wards as the focal points for identifying crime hotspots. To achieve this, we aggregate the crime dataset by date, time, and ward, computing the distinct count of crime incidents (ID) within each ward, thereby obtaining the crime count for that ward. Subsequently, we explore two distinct methodologies to predict hotspots:

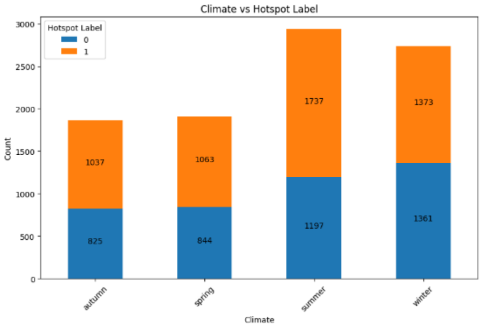
Method 1:

Our approach begins with data aggregation, wherein we group the crime dataset by date, time, and ward, calculating the distinct count of crime incidents within each ward for a given date. This results in the average daily crime count per ward, forming the foundation for subsequent analysis. To provide context for hotspot identification, we compute the daily average crime count for Chicago over the preceding 30 days. This historical reference allows us to assess the current crime count against past trends, aiding in hotspot detection. By merging this data with weather information, we create a comprehensive dataset encompassing temporal, spatial, and meteorological parameters.

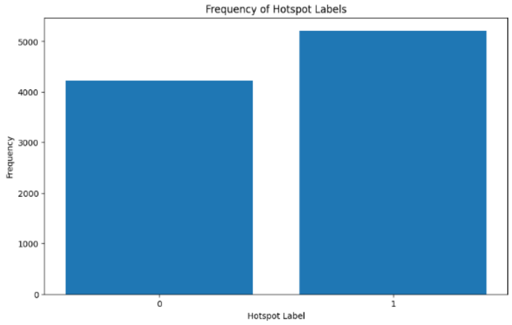


```
+---------------+------------------+
|avg_Crime_count|        prediction|
+---------------+------------------+
|             26| 14.38031587287845|
|             13| 14.38031587287845|
|             11|13.422154735668997|
|              6|13.422154735668997|
|              7|13.422154735668997|
|             23| 13.97268349121281|
|             14| 13.97268349121281|
|             10| 13.97268349121281|
|             29| 13.97268349121281|
|             10| 13.97268349121281|
|              6|  9.88171074399052|
|              5|10.557336459155536|
|              6|10.557336459155536|
|              6|10.557336459155536|
|              9|14.187589796248925|
|             15|14.128878941578572|
|             12|14.128878941578572|
|             19|14.128878941578572|
|              8|14.128878941578572|
|             27|14.128878941578572|
+---------------+------------------+
only showing top 20 rows
```

Now, we construct a Random Forest Regression model to predict the average daily crime count for each ward. Leveraging features such as temperature, precipitation, humidity, and other weather variables, alongside temporal and spatial attributes, we aim to capture the complex dynamics underlying crime occurrences. The model's performance is evaluated using the Root Mean Squared Error (RMSE) metric, which quantifies the deviation between predicted and



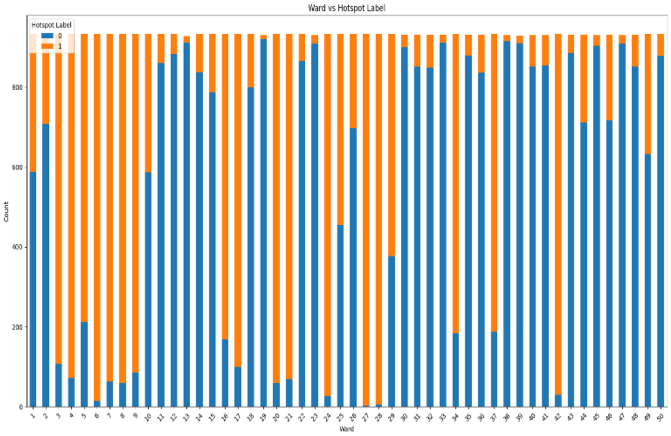
actual crime counts. With an RMSE of 6.5, our regression model demonstrates promising predictive accuracy.



Subsequently, we transition to hotspot identification, employing a classification-based approach. One method involves implementing a function to compare each ward's average crime count with the 30-day historical average for Chicago. Wards exceeding this threshold are labeled as hotspots, while others are classified as non-hotspots. Alternatively, we employ a Random Forest Classifier, which directly predicts hotspot designation based on input features. By incorporating weather variables and past crime counts, the classifier distinguishes between hotspot and non-hotspot wards with a remarkable accuracy of 99.83%.

Method 2:

In Method 2, our approach differs slightly as we assume access to crime counts for each ward. We commence by aggregating the crime dataset, grouping it by date, time, and ward, and computing the distinct count of crime incidents (ID) within each ward for a given date. This yields the average daily crime count per ward, providing valuable insights into localized crime patterns. Utilizing this data, we proceed to determine hotspot labels for each ward.



To accomplish hotspot prediction, we employ a function that calculates the average crime count per ward for a given date, along with the average daily crime count for Chicago over the past 30 days. By comparing each ward's average crime count with the 30-day historical average, we identify wards with crime counts exceeding this threshold as hotspots, assigning them a label of 1.

Conversely, wards with crime counts below the threshold are labeled as non-hotspots (assigned a label of 0). This classification process facilitates the identification of areas with elevated crime rates, enabling targeted intervention strategies.

Upon applying the hotspot labeling function, we obtain a dataframe containing ward-level hotspot labels, providing insights into the spatial distribution of crime hotspots across Chicago. Subsequently, we explore the distribution of hotspot labels across wards over the duration of the dataset. This analysis is visualized through a bar plot, depicting the frequency of hotspot and non-hotspot labels for each ward over the three-year period.

V. Problems Encountered

Special Day Reporting Bias:

Holidays, particularly New Year's Day, show an uptick in crime reports, likely influenced by shift-end reporting. Specifically, there are overwhelming spikes in sexual crimes and crimes involving children. Certain crime types exhibit increases far above the global increase of approximately 80%. Predictive models could erroneously prioritize holidays as high-risk intervals for crime. Data interpretation for holidays may lead to skewed analytical outcomes, impacting the precision of temporal crime predictions.

Bias in Date Recording:

Financial and long-term frauds are uniformly logged on the 1st of each month. This leads to skewed data, misrepresenting actual incident dates. Models may falsely identify the start of the month as a high-risk period. Temporal analysis and prediction accuracy for fraud are consequently compromised.

VI. Conclusion

The **Random Forest Regressor** achieved a root mean squared error (RMSE) of 55.98879793059855 on the test data for predicting crime counts. While this RMSE value provides an overall assessment of the model's performance, a more detailed evaluation can be made by examining the actual versus predicted crime count values shown.

A **RandomForestClassifier** was trained to predict future hotspots and achieved an accuracy of 99.9% on the test set. However, it's important to remember that such a high accuracy might be too good to be true. dataset used to train the model was relatively small, the model might have been able to memorize the specific patterns in the data rather than learning generalizable patterns.

VII. Future Scope

The Impact of External Events: Explore how major events or policy changes, like holidays or legislative shifts, affect crime trends, enhancing predictive accuracy. Also Expanding the dataset over more years for better trend analysis and predictive reliability.

VIII. Citations

Chicago Crime Dataset:

- City of Chicago. "Crimes - 2001 to Present." City of Chicago Data Portal. data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/about_data. Accessed 5 March 2024. (Note: Data analyzed for the period August 6, 2021 to March 2, 2024)

Chicago Weather Dataset:

- Visual Crossing. "Weather data services for Chicago, US [from August 6, 2021 to March 2, 2024]." https://www.visualcrossing.com/. Accessed 05 March 2024.