**Final Paper: Analysis of Relationship Between Weather and Crime in NYC**

**(2013-2019)**

https://github.com/FoundationsOfAnalytics-INFO574/FinalProject\_tianyu\_zhixun\_yuxuan.git

**Yuxuan Fu, Zhixun Guo, Tianyu Zhou**

**Master of Information Systems Management, Washington University in St.Louis**

**FL2024.T81.INFO.574 – Final Paper**

**12/11/2024**

**Data Sources**

This research investigates the relationship between weather conditions and crime patterns in New York City. To conduct the analysis, two primary datasets were used:

1. **Weather Data**: Daily weather information collected from January 1, 2013, to December 31, 2022, was obtained from the National Centers for Environmental Information (NCEI), accessible at [NCEI Website](https://www.ncei.noaa.gov/). The dataset includes variables such as daily precipitation (PRCP), snowfall (SNOW), minimum temperature (TMIN), and temperature delta (TDELTA).

A table with numbers and letters

Description automatically generated

A screenshot of a computer

Description automatically generated

1. **Crime Data**: Crime records from December31, 2005, to December 31, 2019, were sourced from the NYC Open Data portal, accessible at [NYC Open Data](https://opendata.cityofnewyork.us/). The dataset includes various details about arrests, including offense descriptions, demographic information of perpetrators, arrest locations, and time of arrest.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Both datasets were aligned using the date field to ensure a consistent temporal dimension for analysis. This approach enabled a comprehensive exploration of how weather conditions correlate with daily and monthly crime rates over a decade.

**Problem Statement**

The central aim of this research is to investigate whether weather conditions influence crime rates and severity in New York City. Specifically, we seek to answer the following question: "Does weather positively or negatively affect crime patterns in NYC?" Our hypothesis suggests that certain weather conditions, such as temperature fluctuations or extreme precipitation, have a measurable impact on the frequency and severity of crimes. This study aims to explore these dynamics using statistical and machine learning techniques to derive actionable insights.

**Method of Combining Datasets**

To ensure the datasets were compatible for analysis, we aligned the weather and crime datasets using the date field as the key. Given the differing time spans of the two datasets, we selected the overlapping period from January 1, 2013, to December 31, 2019, for the analysis. This ensured both datasets contained synchronized and relevant data for comparison. This process was facilitated by Python, where the Pandas library was utilized for merging operations. The following steps were undertaken:

1. Data Cleaning: Irrelevant and duplicate columns were removed from both datasets, and the remaining data was standardized to ensure consistent formats for date fields.
2. Key Alignment: Both datasets were indexed on the date field to enable precise merging. This ensured that each day’s weather data was accurately paired with the corresponding crime data.
3. Validation: After merging, the dataset was inspected for anomalies, such as duplicate rows or misaligned dates, to ensure the integrity of the combined data.

By merging these datasets within the specified time range, we created a unified dataset that retained the temporal structure necessary for exploring the relationship between weather conditions and crime patterns.

A screenshot of a computer

Description automatically generated

**Dealing with Missing Values and Imbalance**

Handling missing values and class imbalance was an essential step to ensure the reliability of our analysis. The following strategies were employed:

1. **Missing Values**: In the crime dataset, a small fraction of records had missing fields, primarily related to secondary details such as the exact precinct or age group. These records were dropped as they constituted less than 5% of the total dataset. In the weather dataset, missing weather metrics were interpolated linearly to retain continuity.

A screenshot of a computer code

Description automatically generated

1. **Class Imbalance**: To address imbalances in crime types, particularly rare crimes, we employed oversampling techniques for underrepresented classes. Synthetic Minority Oversampling Technique (SMOTE) was utilized to generate synthetic samples for the minority classes, ensuring that the model had adequate representation across all crime types.

These preprocessing steps minimized biases in the analysis, enabling a fair representation of all categories within the datasets.

**Transformations and Interactions**

Several transformations and interaction terms were applied to ensure that the data was suitable for modeling:

1. **Standardization**: Continuous variables, such as precipitation (PRCP), snowfall (SNOW), minimum temperature (TMIN), and temperature delta (TDELTA), were standardized to have a mean of 0 and a standard deviation of This ensured that the variables were on a comparable scale.
2. **Numerical Encoding**: For categorical variables such as borough and offense type, one-hot encoding was employed to represent these variables numerically, ensuring compatibility with machine learning models.

These transformations allowed the models to capture both linear and non-linear relationships effectively, improving the robustness of the analysis.

**Variable Selection**

The selection of variables was critical to ensuring meaningful and interpretable analysis. Key steps included:

1. Relevance to Weather and Crime: Variables were chosen based on their hypothesized relevance to weather and crime patterns. Weather variables included PRCP, SNOW, TMIN, and TDELTA. Crime variables included offense descriptions (e.g., OFNS\_DESC), demographic details (e.g., AGE\_GROUP, PERP\_SEX, PERP\_RACE), and spatiotemporal variables (e.g., ARREST\_BORO, ARREST\_PRECINCT).
2. Temporal Variables: Arrest year, month, and day were included to capture seasonal and daily crime trends.

These selected variables provided a comprehensive basis for understanding the relationship between weather conditions and crime.

**Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) was conducted to understand the distribution and relationships within the data. Key findings include:

1. **Weather Trends**: Seasonal patterns were observed in weather variables, with higher precipitation in spring and lower temperatures in winter. Temperature delta exhibited greater variability during transitional seasons (spring and autumn)

A graph showing a blue line

Description automatically generated

A graph of lines and dots

Description automatically generated

1. **Crime Distribution**: Crime rates were higher in warmer months (May to August), aligning with the hypothesis that temperature positively influences criminal activity.

A graph with a line and a blue line

Description automatically generated

1. **Correlation Analysis**: Weak correlations were found between most weather variables and crime metrics. However, temperature delta showed a mild positive correlation with violent crimes.

A screenshot of a graph

Description automatically generated

These insights guided subsequent modeling by highlighting critical patterns and potential interactions between weather and crime.

**Choosing an Appropriate Technique and Assumptions**

The selection of analytical techniques was based on the characteristics of the data and the research objectives:

1. OLS Regression: Ordinary Least Squares (OLS) regression was chosen to quantify linear relationships between weather variables and crime metrics. Assumptions of linearity, normality of residuals, and homoscedasticity were verified through diagnostic tests.
2. Random Forest: Random Forest was used to capture non-linear relationships and interaction effects between weather variables and crime patterns. The ensemble-based nature of the model reduces overfitting risks while providing feature importance scores.
3. Assumption Validation: For OLS, scatterplots and residual diagnostics were examined to confirm linearity and normality. For Random Forest, sufficient training data was ensured to mitigate overfitting.

These techniques complemented each other, offering both interpretability and predictive power.

**Fitting a Model**

The selected models, OLS Regression and Random Forest, were implemented using Python. The following steps were undertaken to ensure robust model fitting:

1. **Data Splitting**: The dataset was divided into training and testing sets in a 70:30 ratio to validate model performance.
2. **OLS Regression**: The OLS model was fit to the training data using weather variables as predictors. Diagnostic tests were conducted to evaluate residual patterns, ensuring the model met linearity and homoscedasticity assumptions.
3. **Remove Outliers**: Use cooks\_distance to identifty outliers of the data set and Remove with a threshold.

A screenshot of a computer code

Description automatically generated

A graph with numbers and circles

Description automatically generated

1. **Random Forest**: A Random Forest model was trained with 100 estimators, leveraging the scikit-learn library. Hyperparameter tuning was performed using grid search to optimize depth, number of features, and sample splits.
2. **Model Comparison**: Both models were evaluated using metrics such as R-squared and Mean Squared Error (MSE) on the testing set. Feature importance scores from Random Forest provided insights into which variables most significantly impacted crime rates.

A screenshot of a document

Description automatically generated

*OLS on Level of Crime vs Weather*

*A screenshot of a computer

Description automatically generated*

*Random Forest Classifier on Level of Crime vs Weather*

*A screenshot of a computer

Description automatically generated*

*Logistic Regression on Level Crime vs Weather*

*A screenshot of a computer

Description automatically generated*

*OLS on Num Crime vs Weather*

**Evaluating the Model and Overall Modeling**

Model evaluation and insights were derived through comprehensive analysis:

1. **OLS Regression Evaluation**:
   * **R-squared**: The OLS model explained only a small portion of the variance in crime metrics, indicating limited predictive power.
   * **Residual Analysis**: Residual plots revealed non-random patterns, suggesting that linear assumptions might not fully capture the relationship.

A graph of a plot with blue dots

Description automatically generated with medium confidence

1. **Random Forest Evaluation**:
   * **Accuracy**: Random Forest achieved higher predictive accuracy compared to OLS.
   * **Feature Importance**: Temperature delta (TDELTA) and precipitation (PRCP) emerged as the most influential variables.
2. **Comparison**:
   * OLS regression provided insights into linear relationships, but its assumptions limited its application to more complex patterns.
   * Random Forest effectively captured non-linear relationships and interactions, offering better generalization and interpretability for decision-making.

**Overall Insights**: The modeling results underscore the importance of employing diverse techniques to capture complex dynamics. Random Forest's ability to reveal variable importance provides actionable insights, particularly regarding how weather fluctuations influence crime.

**Conclusions and Reasoning**

The study explored the relationship between weather conditions and crime patterns in New York City using both statistical and machine learning methods. Key conclusions include:

1. **Weather's Influence on Crime**:
   * Temperature fluctuations and precipitation were found to be significant predictors of crime rates, particularly violent crimes.
   * Seasonal patterns indicated higher crime rates during warmer months.
2. **Model Effectiveness**:
   * Random Forest outperformed OLS regression in capturing non-linear relationships and providing insights into variable importance.
   * Feature importance analysis highlighted the pivotal role of temperature delta and precipitation.
3. **Practical Implications**:
   * Insights from the analysis can inform law enforcement strategies, such as allocating resources based on weather forecasts.
   * The findings underscore the value of integrating weather data into crime prevention frameworks.

**Recommendations for Future Research**:

* Incorporate additional variables such as socioeconomic factors to enhance predictive power.
* Explore real-time applications of the models for dynamic resource allocation.

**References**

1. National Centers for Environmental Information (NCEI). "Daily Weather Data." <https://www.ncei.noaa.gov/>
2. NYC Open Data. "NYPD Arrest Data." <https://opendata.cityofnewyork.us/>
3. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
4. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.