Crime Forecasting Model Report

Overview of the Model

The SARIMA crime forecasting model was developed to predict monthly violent crime counts in Charlotte, NC, using historical data from 2017 to 2024. The primary goal of the model is to analyze crime trends and forecast crime counts for the next year (2024-2025) to aid decision-making by the Charlotte-Mecklenburg Police Department (CMPD) and other stakeholders.

The model utilizes **seasonality**, **trend analysis**, and **historical patterns** to provide actionable insights. It was optimized iteratively to improve accuracy and performance metrics, ensuring it meets the project's requirements.

Model Results and Insights

1. **Performance Metrics**:
    - **R-squared**: 0.30
        - Indicates the model explains 30% of the variance in the data, reflecting reasonable predictive power for a noisy and complex dataset like crime counts.
    - **Mean Absolute Error (MAE)**: 78.37
        - The average prediction error suggests the model's forecasted monthly crime counts deviate by approximately 78 crimes from the actual values.
    - **Root Mean Squared Error (RMSE)**: 99.91
        - Highlights the typical magnitude of error, penalizing larger deviations more heavily than MAE.
2. **Residual Analysis**:
    - **Random Distribution**: Residuals showed no discernible patterns, indicating the model captured key trends and seasonality effectively.
    - **Autocorrelation**: The Ljung-Box test confirmed no significant autocorrelation, validating the model's ability to capture temporal dependencies.
    - **Constant Variance**: Residuals exhibited consistent spread, ensuring stable predictions across the dataset.
    - **Unexplained Variance**: A residual variance of 70% suggests room for improvement through additional explanatory variables.
3. **Strengths**:
    - **Seasonality**: The model captures recurring patterns, such as monthly crime surges, particularly around specific periods like summer and holidays.
    - **Trend Prediction**: Effectively forecasts long-term increases or decreases in crime trends.

    - **Validated Predictions**: Residual analysis confirmed the model's reliability for short-term forecasting.
4. **Weaknesses**:
    - **Limited Explanatory Power**: R-squared of 0.30 indicates significant variability in crime counts remains unexplained.
    - **Sensitivity to Noise**: The model struggles with unpredictable short-term crime spikes or dips.
    - **Lack of External Variables**: Excluding factors like holidays, socioeconomic data, and policing initiatives reduces the model's accuracy.

CMPD Applications

The model provides CMPD and other stakeholders with several practical applications:

1. **Resource Allocation**:
    - Proactively deploy resources to areas or times with expected high crime counts.
    - Optimize staffing schedules and prioritize high-risk periods, such as holidays.
2. **Strategic Planning**:
    - Use forecast trends to design long-term crime-prevention strategies.
    - Coordinate with city departments to reduce environmental opportunities for crime.
3. **Community Engagement**:
    - Share forecasted trends with community organizations to encourage collaboration on crime-reduction programs.
    - Raise public awareness about high-risk periods and preventive measures.
4. **Policy Evaluation**:
    - Compare actual crime counts against forecasts to measure the effectiveness of new crime policies or interventions.
5. **Data-Driven Budgeting**:
    - Justify resource allocation for crime prevention programs based on forecasted trends.

Model Optimization Process

1. **Initial Model**:
    - **Parameters**: SARIMA(2, 1, 0) x (0, 1, 1, 12).
    - **Performance Metrics**:
        - R-squared: 0.21
        - MAE: 73.42
        - RMSE: 105.76

- **Rationale for Initial Parameters**:
  The initial parameters were manually selected based on general SARIMA guidelines:
  - `order=(2, 1, 0)` assumes a lag of 2 for autoregression (`p`), first-order differencing (`d`), and no moving average (`q=0`).
  - `seasonal_order=(0, 1, 1, 12)` captures monthly seasonality (`S=12`) with first-order seasonal differencing (`D=1`) and a seasonal moving average component (`Q=1`).
- **Weaknesses of Initial Model**:
  - The R-squared value of 0.21 indicated the model was not capturing much of the variance in the data.
  - Residual analysis revealed some patterns that suggested suboptimal parameter choices, potentially underfitting the data.

---

2. **Optimized Model (First Iteration)**:
   - **Changes Made**:
     - Implemented **grid search** to automate the selection of SARIMA parameters:
       - Searched combinations of `p`, `d`, `q` for non-seasonal terms and `P`, `D`, `Q`, `s` for seasonal terms.
       - Focused on reasonable ranges for parameters to avoid overfitting:
         - `p, d, q` in {0, 1, 2}.
         - `P, D, Q` in {0, 1, 2}.
         - Seasonal period `S=12` (monthly data).
     - Identified the best-performing parameters:
       - **Non-seasonal order**: `(2, 1, 1)`
       - **Seasonal order**: `(2, 1, 0, 12)`
   - **Rationale for Parameter Changes**:
     - Adding a moving average term (`q=1`) in the non-seasonal order helped smooth out short-term fluctuations.
     - Introducing an additional seasonal autoregressive component (`P=2`) improved the model's ability to capture repeating patterns in crime data.
   - **Performance Improvements**:
     - **R-squared**: Improved from 0.21 to 0.30 (+42.9% improvement).
     - **RMSE**: Reduced from 105.76 to 99.91 (-5.5% improvement).
     - **MAE**: Increased slightly from 73.42 to 78.37, reflecting a trade-off between reducing larger errors (as seen in RMSE) and smaller absolute deviations.
   - **Insights**:

        - The optimization process demonstrated that parameter tuning can significantly improve the model's ability to explain variance and capture seasonal patterns.
        - While MAE increased slightly, the model performed better overall due to reduced RMSE and higher R-squared.

---

3. **Residual Analysis and Validation**:
    - **Changes Made**:
        - Conducted residual analysis to validate the optimized model:
            - Checked for random distribution of residuals, confirming no systematic patterns were left unexplained.
            - Used the Ljung-Box test and autocorrelation plots to ensure residuals were not autocorrelated.
            - Verified homoscedasticity (constant variance) in residuals over time.
    - **Rationale for Residual Analysis**:
        - Ensuring residuals met the assumptions of randomness, independence, and constant variance validated the model's fit.
        - Residual diagnostics highlighted the absence of significant patterns, suggesting the optimized model captured trends and seasonality effectively.
    - **Results**:
        - Randomly distributed residuals and no significant autocorrelation validated the optimized parameters.
        - The Ljung-Box p-value of 0.37 confirmed that the residuals were not correlated.

---

4. **Final Model Metrics**:
    - **R-squared**: 0.30 (+42.9% improvement over the initial model).
    - **MAE**: 78.37 (+6.7% from the initial model).
    - **RMSE**: 99.91 (-5.5% improvement over the initial model).

Summary of Optimization Changes

Aspect	Initial Model	Optimized Model	Change
Non-Seasonal Order	`(2, 1, 0)`	`(2, 1, 1)`	Added moving average term (`q=1`)
Seasonal Order	`(0, 1, 1, 12)`	`(2, 1, 0, 12)`	Added seasonal autoregressive term (`P=2`)
R-squared	0.21	0.30	+42.9% improvement
RMSE	105.76	99.91	-5.5% improvement
MAE	73.42	78.37	Slight increase (trade-off for RMSE)
Residual Diagnostics	Minimal	Extensive	Validated randomness, no autocorrelation

---

Recommendations for Future Improvements

1. **Include External Factors**:
   - Incorporate variables such as holidays, weather data, unemployment rates, or major city events to improve explanatory power.
2. **Explore Alternative Models**:
   - Test machine learning approaches like **Prophet**, **XGBoost**, or **LSTMs**, which handle non-linear relationships and changing trends better.
3. **Refine Data Granularity**:
   - Experiment with weekly aggregation instead of monthly counts to balance noise reduction and resolution.
4. **Real-Time Forecasting**:
   - Develop a real-time pipeline to integrate updated crime reports into forecasts for dynamic decision-making.

Conclusion

The SARIMA model is a robust tool for forecasting crime counts in Charlotte, NC, providing meaningful insights into long-term trends and seasonal patterns. While it is well-suited for resource planning and strategic decision-making, integrating additional data and experimenting with advanced models could further enhance its predictive accuracy and utility for CMPD and other stakeholders.