Crime in Tucson: Predicting Crime Severity Using Demographics, Time, and Location

Authored by Ashutosh Dayal, Jakob Garcia, Pri Vaghela

1. Introduction

For law enforcement and the communities they protect, it is essential to comprehend the seriousness of crimes, whether they are felonies or misdemeanors. Authorities can more efficiently deploy resources, carry out regulations, and make areas safer when they are aware of the variables that affect the intensity of crimes. In order to anticipate the severity of crimes in Tucson, Arizona, this research looks at how location, timing, and demographics interact to influence criminal behavior.

We accomplish this by combining neighborhood demographics, crime statistics over time, and data from the Tucson Police Department. By combining these datasets, we hope to find trends that help explain why serious crimes are more likely to occur in particular locations or during particular times. We train algorithms to categorize crimes as either felonies or misdemeanors using machine learning models like Random Forest and Logistic Regression. By highlighting patterns that may not be immediately apparent, exploratory visualizations aid in our understanding of the context behind the data.

We keep the ethical ramifications in mind when we construct these models. Certain communities may be disproportionately impacted by predictive algorithms that serve to reinforce pre-existing biases in the data. In order to tackle this, we prioritize equity at every stage of the process by making sure that the models are created and used in a transparent and equitable manner.

This initiative is fundamentally about more than just statistics and forecasts. It's about making strategic use of data to help make better decisions and build more equitable and safe communities. We believe that by fusing state-of-the-art methods with a dedication to moral behavior, this work offers insightful information to scholars and decision-makers.

---

2. Related Work

Previous studies have shown strong associations between crime patterns and variables such as demographics, socioeconomic status, and geographic location.

Crime prediction research often highlights the role of socioeconomic factors in influencing criminal behavior. Yildiz, Ocal, and Yildirim (2013) investigated the effects of income, education, and unemployment on crime rates using individual-level panel data from Turkey. Their findings underscore the strong inverse relationship between income and crime, emphasizing the deterrent effect of higher wages. Similarly, education was found to influence crime indirectly by increasing opportunity costs, although the effect of low education on crime was significantly stronger than that of higher education. Unemployment, however, exhibited only a marginal impact, largely due to alternative income sources for unemployed individuals, such as unregistered employment or government aid (Yildiz et al., 2013). These insights align with the objectives of this project, reinforcing the importance of including demographic and socioeconomic variables in predictive models for crime severity. Furthermore, their application of dynamic panel data modeling to address biases in crime data informs the methodological approach employed in this project, ensuring robust and unbiased predictions.

While predictive modeling offers promise in crime analysis, it also faces significant ethical challenges. Susser (2019) critiques predictive policing tools for perpetuating systemic inequalities and embedding racial and socioeconomic biases from historical data. He emphasizes the dangers of feedback loops, where predictions based on biased data lead to over-policing certain communities, reinforcing existing disparities. Susser also raises concerns about the dehumanizing aspect of algorithmic profiling, which undermines the presumption of innocence and moral agency. Importantly, he suggests alternative contexts for using predictive tools, such as in social services, to prioritize individual welfare over punitive measures (Susser, 2019). These critiques inform the ethical considerations of this project, highlighting the need for transparency and fairness in predictive modeling.

Predictive modeling in this domain has been explored, yet it often encounters challenges, including biased outcomes due to historical data limitations. This project advances these efforts by integrating diverse datasets and refining the predictive approach, focusing on ethical implementation. Through careful selection and evaluation of features, this study aims to balance the utility of prediction with minimizing potential ethical risks.

---

3. Methods

The methods employed in this project are designed to address the primary objective: predicting the severity of crimes (misdemeanor or felony) in Tucson using demographic, temporal, and neighborhood-based location data. This section details the datasets, preprocessing steps, exploratory analyses, and modeling techniques used to develop and evaluate the predictive models.

3.1 Data Sources

The analysis relies on a combination of datasets from the Tucson Police Department and publicly available neighborhood demographic data:

1. **Tucson Police Arrest Data (TPA) (Years 2019-2021)**:
    - Contains details about arrests, including age, race, sex, date, time, and type of charge (misdemeanor or felony).

2. **Tucson Police Reported Crimes Data (TPRC)**:
    - Provides detailed crime reports categorized by type, time of occurrence, and geographic location.
3. **Neighborhood Demographics**:

    - Includes data on population age distribution, income, education levels, employment, and racial composition for Tucson neighborhoods.

These datasets are sourced from the City of Tucson Data Hub and merged to form a unified framework for analysis.

3.2 Data Preprocessing

3.2.1 **Data Cleaning**

- Removed missing and inconsistent entries to ensure data integrity.
- Standardized column formats across datasets to enable smooth merging.
- Filtered data to include only records relevant to the target prediction (misdemeanor vs. felony).

3.2.2 **Data Merging**

- Neighborhood demographic data was mapped to arrest records using geographic identifiers, such as neighborhood names or IDs.
- Temporal variables (e.g., time and date of crime) were combined with demographic and arrest data for a comprehensive dataset.

3.2.3 **Feature Engineering**

- **Cyclical Encoding of Temporal Variables**:
  Transformed time-related variables, such as day of the week and hour of occurrence, into sinusoidal features to capture cyclical patterns in crime activity.
- **Standardization**:
  Demographic variables, such as income and education, were scaled to ensure consistent interpretation and performance in machine learning models.
- **One Hot Encoding**:
  Converted categorical variables into dummy numeric variables for compatibility with

machine learning algorithms and to mitigate the risk of a wrongful interpretation of a numerical significance.

3.3 Exploratory Data Analysis (EDA)

EDA provided valuable insights into the relationships between variables and the target outcome:

- **Crime Frequency by Type**:
  Analyzed the distribution of crimes (e.g., assault, robbery) to understand broader trends in criminal activity.

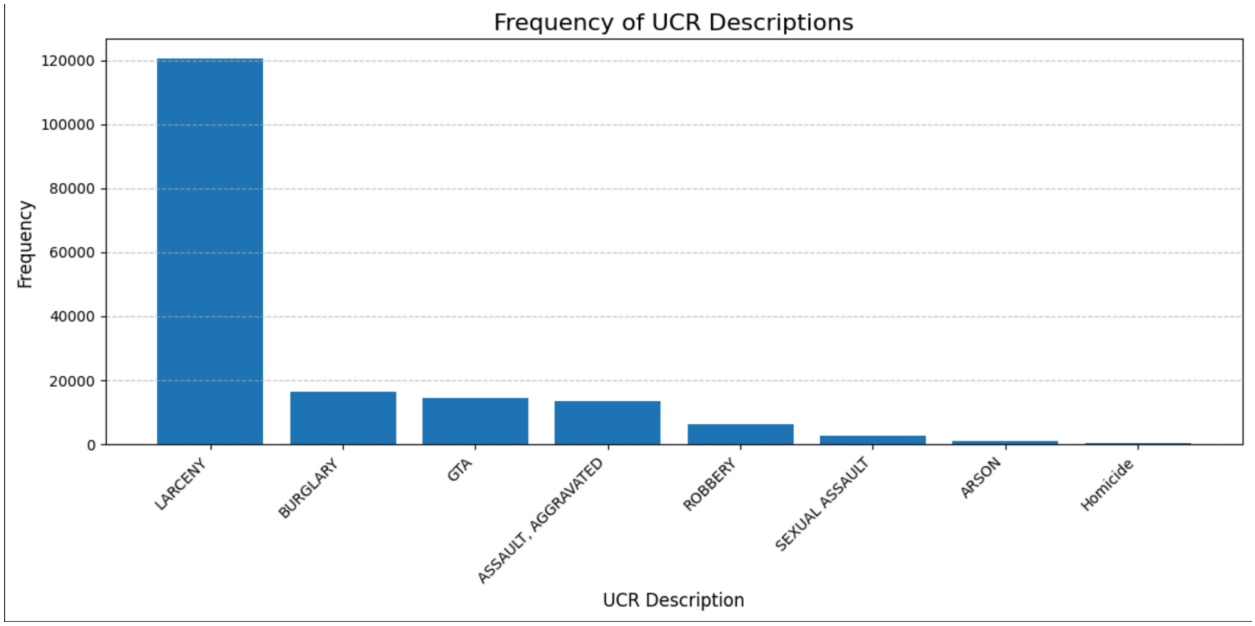

Fig 1:Distribution of crime types based on UCR (Uniform Crime Reporting) descriptions, highlighting larceny as the most frequent offense.

- **Demographic Correlations**:
  Explored the influence of socioeconomic factors, such as income levels and education

attainment, on crime severity.

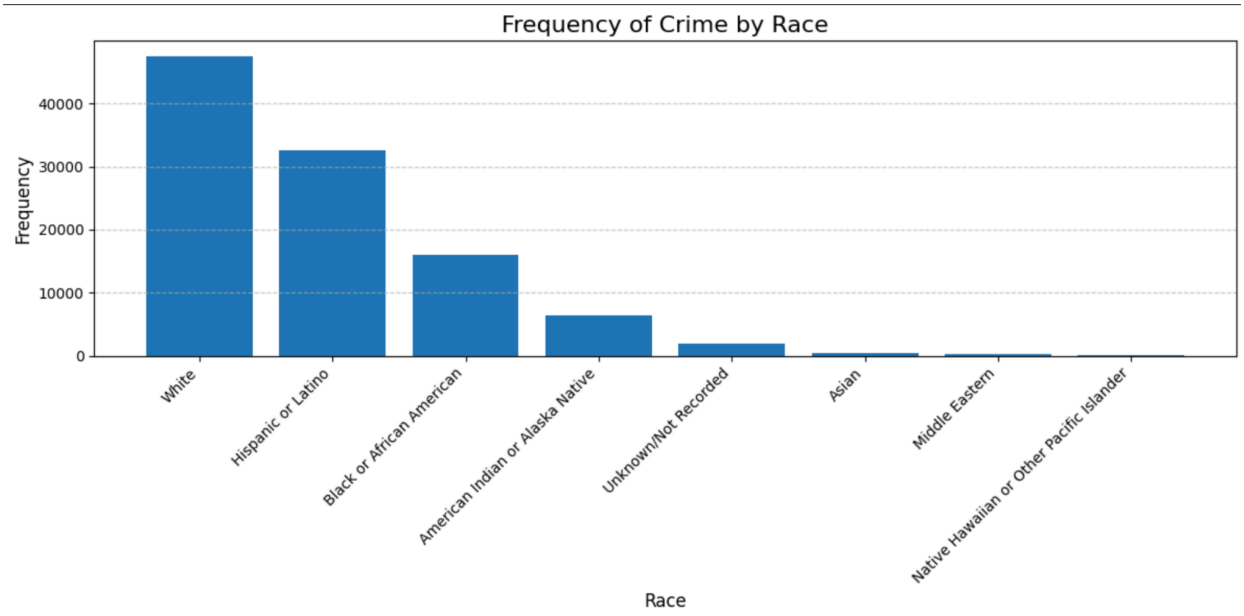

Fig 2:Frequency of crime occurrences by race, with the highest counts observed among White and Hispanic or Latino individuals.

- **Temporal Trends**:

  Visualized crime occurrences across different times of day and days of the week to identify patterns.

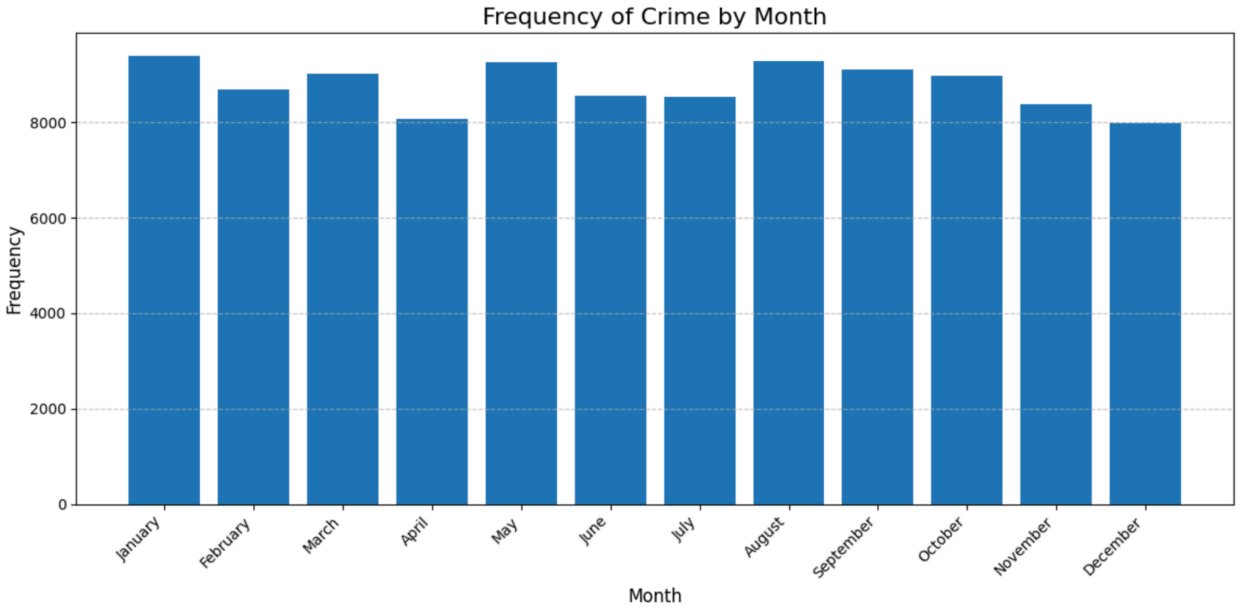

Fig 3:Monthly distribution of crime frequency, showing relatively consistent rates throughout the year with slight variations.

While these visualizations helped contextualize the data, the primary focus remained on predictive modeling.

3.4 Modeling

3.4.1 **Random Forest Classifier**

- **Rationale**:
  Chosen for its ability to handle non-linear relationships and automatically determine feature importance.
- **Features**:
  Combined demographic, temporal, and geographic variables.
- **Hyperparameter Tuning**:
  Used grid search to optimize parameters, such as the number of trees and maximum depth.

3.4.2 **Logistic Regression**

- **Rationale**:
  Provides a simple, interpretable model to identify key predictors of crime severity.
- **Features**:
  Same as the Random Forest, ensuring a consistent comparison.
- **Cross Validation**:
  Performed 10-fold cross validation and hyperparameter tuning to determine the best inverse of regularization strength to prevent overfitting

3.5 Evaluation Framework

To assess model performance and ensure robustness:

- The dataset was split into training (80%) and testing (20%) subsets.

- Both models were evaluated using metrics such as:
    - **Accuracy**: Proportion of correctly classified cases.
    - **Precision**: Ability to correctly identify felonies without including misdemeanors.
    - **Recall**: Sensitivity in identifying all felonies.
    - **F1-Score**: Balances precision and recall for a holistic evaluation.
- Cross-validation was employed to minimize overfitting and validate model generalizability.

3.6 Key Considerations

Ethical considerations informed every step of the methodology:

- Efforts were made to minimize bias in the data, particularly in demographic and geographic variables.
- The transparency of preprocessing and modeling decisions was prioritized to ensure replicability and accountability.

By integrating rigorous data preparation with thoughtful modeling and evaluation, this methodological framework ensures reliable and actionable predictions while addressing the complexities of crime severity in Tucson.

---

4. Results

4.1 Model Performance

- **Random Forest Classifier**:
    - Accuracy: 87.30%
    - Precision: 87%
    - Recall: 87%
    - F1-Score: 87%

- Insight: The most important features were related to the type of crime, and the type of arrest made. Interestingly, time information about the time of arrest seemed to be important for determining the severity as well.

- **Logistic Regression**:
  - Accuracy: 87.64%
  - Precision: 88%
  - Recall: 88%
  - F1-Score: 88%
  - Insight: With misdemeanor as the class assigned to 1 and felony as the class assigned to 0, the odds ratios for the coefficients of the logistic regression were calculated and assessed. It was found that Cases classified under UCR_06 (larceny) are 2.34 times more likely to be misdemeanors than felonies. Arrests of type FR (presumably field release) are 2.2 times more likely to result in misdemeanor classification. Arrests of type BK (presumably booking) are 44% less likely to be classified as misdemeanors compared to the baseline. Web-reported calls are associated with a 35% lower likelihood of misdemeanor classification.

	Odds ratio
UCR_06	2.339459
arr_type_FR	2.199953
CallSource_Officer-Initiated	1.176941
arr_type_PT	1.170896
age	1.160464

	Odds ratio
UCR_05	0.717782
arr_type_S	0.687192
Year_2021	0.684515
CallSource_Web Reported	0.645249
arr_type_BK	0.558505

Table 1 and 2: Top 5 highest and top 5 lowest odds ratios based on the coefficients of the logistic regression model, with UCR_06 (crime category larceny) and arrest type FR being the strongest predictors of a misdemeanor classification. Whereas, features like arrest type BK and web reported calls reduce the likelihood of a misdemeanor classification.

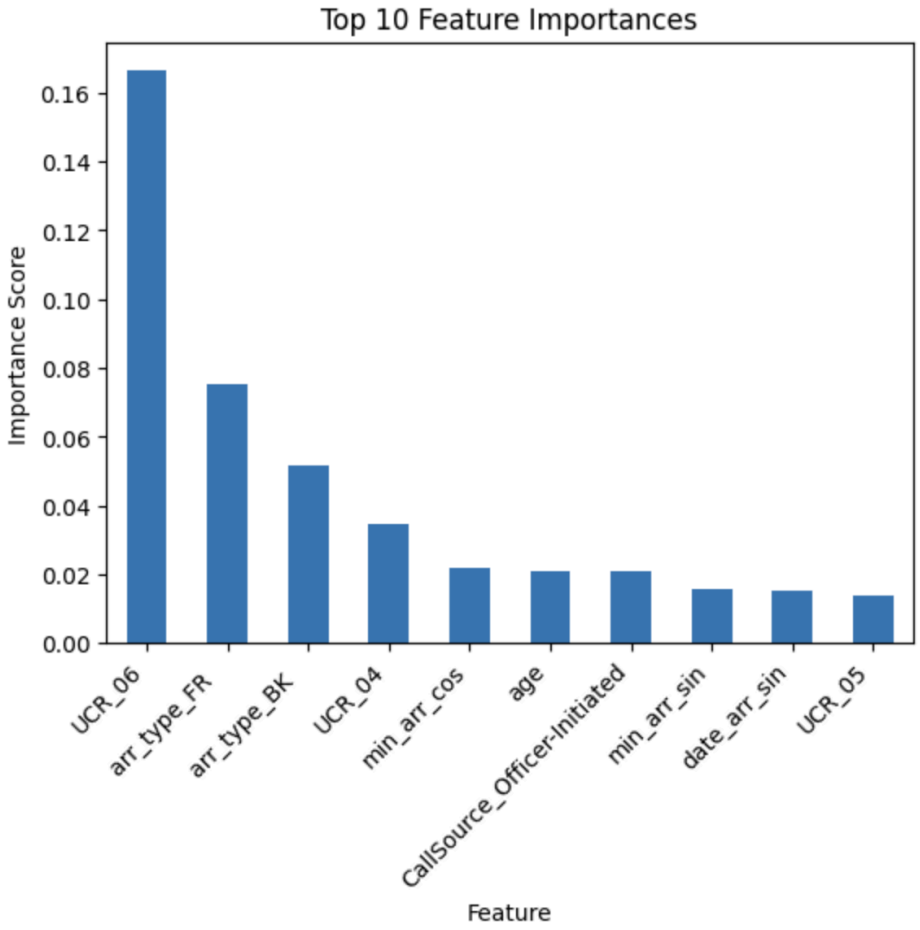

Fig 4:Top 10 most important features identified by the Random Forest model, with UCR_06 (crime category larceny) and arrest types (FR and BK) being the most influential predictors of crime severity.

4.2 Key Findings

- The most important crimes for determining severity were larceny, assault, and burglary according to the random forest model. This makes sense since larceny is more often a misdemeanor than anything else, and things like assault/aggravated assault and burglary are more often felonies. The type of arrest made was also important. The two most significant types of arrests were FR (presumably field release) and BK (presumably

booking). A field release is when an officer issues a citation and releases the individual at the scene instead of transporting them to a detention facility. This is common for lower-level offenses. A booking indicates an arrest where the individual was booked into jail or a detention facility. These two types of arrests clearly paint the picture of the severity of the crime.

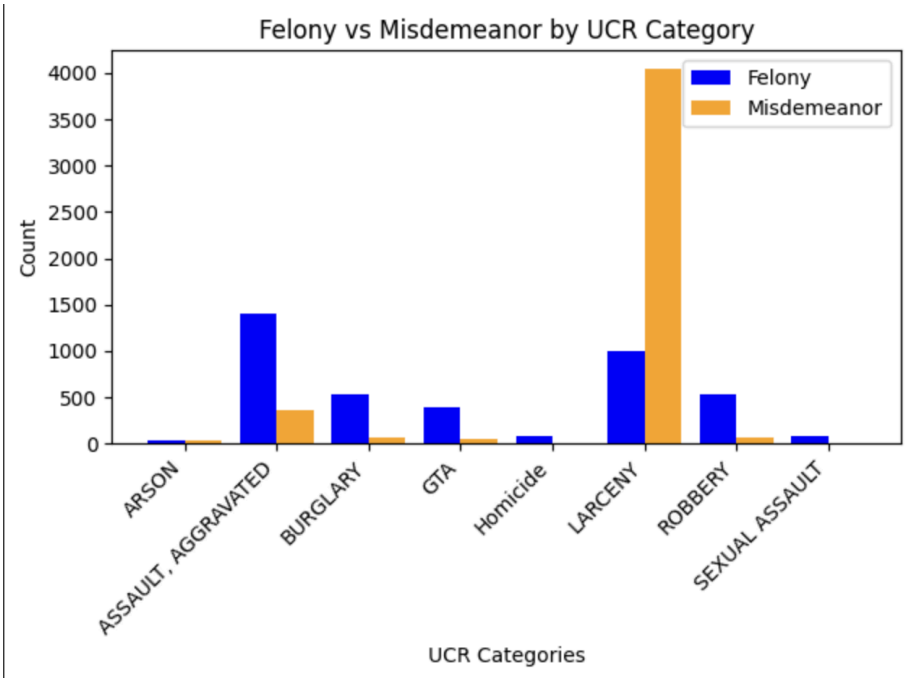

Fig 5:Comparison of felonies and misdemeanors across UCR categories, highlighting larceny's dominance among misdemeanors and aggravated assault's prevalence among felonies.

- Temporal features such as time of day and day of the week were strong predictors of crime severity. The occurrence of arrests by minute was also a key feature, visualized below. There is a probability of the data being inaccurate due to the police officers rounding off the minute arrests were made when filing their reports, leading to the 0^{th} minute having the most arrests.

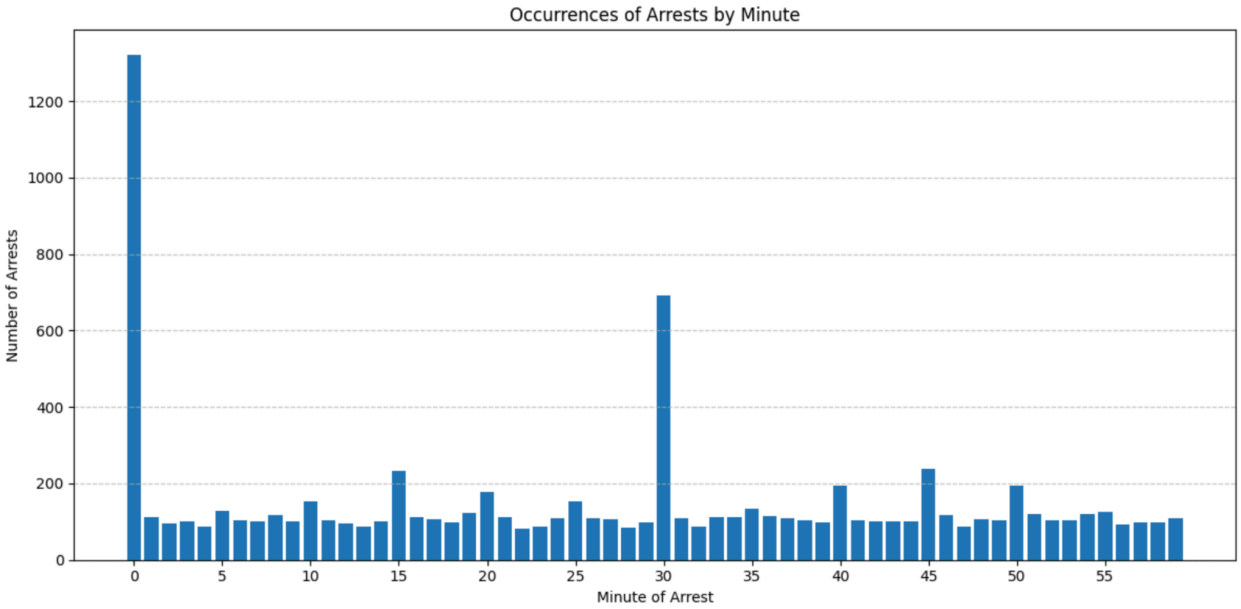

Fig 6:Distribution of arrests by the minute, highlighting peaks at specific intervals, with the model identifying the minute of arrest as a significant predictor of crime severity.

- The odds ratios of the coefficients of the logistic regression model shows that crimes categorized as UCR_06 (larceny) and arrests of type FR (field report) are strongly associated with misdemeanor classification. This ties into what was found from the random forest model. Older individuals also have a slightly higher likelihood of misdemeanor charges. On the other hand, web-reported calls and arrests of type BK (booking) are more likely to result in felony charges, indicating potential differences in case handling. The temporal trend in 2021 suggests a contextual factor that reduced misdemeanor classifications. Future work can be done to determine what in 2021 specifically caused higher odds of a crime resulting in a felony.

---

5. Conclusion

5.1 Summary

This project successfully demonstrated that crime severity in Tucson can be predicted using a combination of demographic, temporal, and geographic data. Both the Random Forest and

Logistic Regression models provided valuable insights, with Logistic Regression achieving slightly higher accuracy (87.64%) and interpretability, while Random Forest excelled in feature importance analysis. Key predictors of severity included crime type (e.g., larceny, burglary, aggravated assault), arrest types (field release or booking), and temporal features such as the minute of arrest. The latter, while significant, revealed potential data recording biases, which were accounted for in the analysis.

These findings underscore the importance of integrating diverse datasets and advanced machine learning techniques to better understand and predict criminal behavior, ultimately supporting law enforcement and public safety efforts.

5.2 Ethical Considerations

- **Bias in Data**: The study highlighted the risks of perpetuating historical biases in predictive policing, emphasizing the need for fairness and transparency in both data preparation and model implementation.
- **Responsible Use**: Predictive tools should enhance decision-making rather than replace human judgment, ensuring equitable outcomes and avoiding over-reliance on algorithmic outputs.

5.3 Future Work

- **Dynamic Predictions**: Incorporating real-time data, such as live crime reports or updated demographic information, to improve model adaptability and accuracy.
- **Causal Analysis**: Exploring the impact of interventions, such as increased street lighting or enhanced community programs, on reducing crime severity.
- **Addressing Data Limitations**: Investigating potential biases in reported data (e.g., the tendency to round timestamps) and their implications for predictive accuracy.

By addressing these areas, future research can build on this study's findings to develop even more robust, equitable, and actionable frameworks for crime prediction and prevention.

6. References

City of Tucson GIS. (2019). *Neighborhood age demographics*. Retrieved from https://gisdata.tucsonaz.gov/datasets/cotgis::neighborhood-age-demographics/explore?showTable=true

City of Tucson GIS. (2019). *Neighborhood educational attainment*. Retrieved from https://gisdata.tucsonaz.gov/datasets/cotgis::neighborhood-educational-attainment/explore?showTable=true

City of Tucson GIS. (2020.). *Neighborhood employment demographics*. Retrieved from https://gisdata.tucsonaz.gov/datasets/cotgis::neighborhood-employment-demographics/explore?location=32.197863%2C-110.889177%2C10.12&showTable=true

City of Tucson GIS. (2019). *Neighborhood income*. Retrieved from https://gisdata.tucsonaz.gov/datasets/cotgis::neighborhood-income/explore?showTable=true

City of Tucson GIS. (2019). *Neighborhood race demographics*. Retrieved from https://gisdata.tucsonaz.gov/datasets/cotgis::neighborhood-race-demographics/explore?showTable=true

City of Tucson GIS. (2024). *Tucson police reported crimes*. Retrieved from https://gisdata.tucsonaz.gov/datasets/cotgis::tucson-police-reported-crimes/explore?showTable=true

City of Tucson GIS. (2019, upd. 2023). *Tucson police arrests 2019 open data*. Retrieved from https://gisdata.tucsonaz.gov/datasets/cotgis::tucson-police-arrests-2019-open-data/explore?location=0.000240%2C-111.225550%2C0.00&showTable=true

City of Tucson GIS. (2020, upd. 2023). *Tucson police arrests 2020 open data*. Retrieved from https://gisdata.tucsonaz.gov/datasets/cotgis::tucson-police-arrests-2020-open-data/explore?location=0.012819%2C-111.225550%2C0.00&showTable=true

City of Tucson GIS. (2021, upd. 2023). *Tucson police arrests 2021 open data*. Retrieved from https://gisdata.tucsonaz.gov/datasets/cotgis::tucson-police-arrests-2021-open-data/explore?location=0.683544%2C-111.225550%2C0.00&showTable=true

Yildiz, R., Ocal, O., & Yildirim, E. (2013). The effects of unemployment, income and education on crime: Evidence from individual data. Journal of Economic & Management Perspectives, 7(2), 32.

Susser, D. (2021). Predictive policing and the ethics of preemption. The ethics of policing: New perspectives on law enforcement, 268-292.