MAP 6114 HOMICIDE PREDICTION TECHNICAL REPORT

Sara Basili
MAP 6114
University of West Florida
Pensacola, Florida
seb88@students.uwf.edu

1 Problem Definition

According to the Metropolitan Crime Commission, New Orleans has been recently declared the murder capital of the United States. The main objective of this project is to predict the likelihood of a homicide happening in the greater area of New Orleans, based on the electronic police reports filed in the year 2022. Being able to better understand the probability of a homicide happening in a specific district could bring insights about the spatial and temporal patterns of violent crime, including the seasonality, clustering, and distribution of homicide occurrences. The project aims to develop a machine learning model that can predict the likelihood of a homicide based on various demographic factors as noted in the police reports. The algorithms will be trained on a subset of the data, and then tested on the remaining data to evaluate their accuracy in predicting homicides. The model could provide significant findings for policymakers and law enforcement agencies to better understand the homicide patterns in various areas across the Orleans parish. The report includes the methodology and results of the analysis, as well as the ethical considerations and potential biases linked with this approach to predictive modeling. The data utilized for this project is a public dataset retrieved from the open data portal for the city of New Orleans.

A link to this portal is provided here: https://datadriven.nola.gov/home/

The analysis is based on the Electronic Police Reports for the year 2022 in the city of New Orleans. This includes all police reports filed by New Orleans Police Department (NOPD) officers, containing incident and supplemental reports. The dataset contains approximately 90,000 observations and 23 features. Each police report is described by an item number, district, location, disposition, signal, charges, offender race, offender gender, offender age, victim age, victim gender, and victim race. This dataset may change dynamically due to Police Reports being updated when subsequent information is determined as a result of an investigation.

The following disclaimer has been provided by the NOPD through the City of New Orleans website:

"The New Orleans Police Department does not guarantee (either expressed or implied) the accuracy, completeness, timeliness, or correct sequencing of the information. The New Orleans Police Department will not be responsible for any error or omission, or for the use of, or the results obtained from the use of this information. For instance, the data contains ages that may be negative due to data entry errors. NOPD has chosen to publish the data as it exists in the source systems for transparency and has instituted data validation where appropriate to ensure quality data in the future. All data visualizations on maps should be considered approximate and attempts to derive specific addresses are strictly prohibited. The New Orleans Police Department is not responsible for the content of any off-site pages that are referenced by or that reference this web page other than an official City of New Orleans or New Orleans Police Department web page. The user specifically acknowledges that the New Orleans Police Department is not responsible for any defamatory, offensive, misleading, or illegal conduct of other users, links, or third parties and that the risk of injury from the foregoing rests entirely with the user. Any use of the information for commercial purposes is strictly prohibited. The unauthorized use of the words "New Orleans Police Department," "NOPD," or any colorable imitation of these words or the unauthorized use of the New Orleans Police Department logo is unlawful. This web page does not, in any way, authorize such use."

To evaluate the performance of our models in predicting the likelihood of a homicide, several performance metrics will be utilized. These metrics will provide insights into the accuracy, precision, recall, and overall performance of each model.

Accuracy: This metric represents the proportion of correct predictions out of all predictions made by the model. While accuracy is a simple and easy-to-understand metric, it may not be appropriate when dealing with imbalanced datasets, as it can be misleading in scenarios where the majority class dominates the data.

Precision: Precision quantifies the proportion of true positive predictions out of all positive predictions made by the model. It is a useful metric when the cost of false positives is high, as it measures the ability of the model to correctly identify the positive (in this case, homicide) instances.

Recall: Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions out of all actual positive instances. This metric is particularly important when the cost of false negatives is high, as it evaluates the ability of the model to identify all relevant instances.

F1 Score: The F1 score is the harmonic mean of precision and recall, providing a single value that balances both metrics. It is especially useful when dealing with imbalanced datasets or when there is an unequal importance between false positives and false negatives.

1.1 Ethical Concerns and Biases

When using machine learning models to predict crime data, several ethical problems and biases may arise. The first concern involves possible biases in the training data. If the training data utilized to train the machine model is biased, the model itself will also be biased. The data often only focuses on offenses generally associated with specific demographic groups and neighborhoods, crimes known as "street crimes" such as theft and carjackings. Domestic violence and white-collar crimes, such as fraud and embezzlement, on the other hand, tend to receive less attention and are not as reported. Overrepresentation or underrepresentation of a specific category in the data may lead to inconsistent results and an inaccurate representation of the world. If certain neighborhoods or demographic groups, for example, are overrepresented in the training data, the model may be more likely to predict crimes among those neighborhoods or racial/ethnic groups resulting in possible discrimination. This could sustain stereotypes and reinforce systemic biases in the criminal justice system, which could lead to negative consequences for marginalized communities. If a group of people begins to feel unfairly targeted by law enforcement based on such predictions, they may be less likely to engage in normal activities, resulting in a further marginalization of a specific community.

The data may also be subject to error, as it could have been entered into the system incorrectly by law enforcers or it could have been overlooked. Another major concern involves self-fulfilling prophecies when resulting predictions are acted upon. If law enforcement officers are directed to increment surveillance in a specific area rather than another, the increased presence of police could lead to more arrests, which could then be used to justify the accuracy of the predictive model.

Predictive models on crime rates could further lead to preemptive punishments, in which individuals are punished based solely on their predicted likelihood of committing a crime as opposed to a crime they have actually committed.

It is important to keep into consideration these problems when building a predictive model using machine learning and observing its results. Taking into consideration all of these factors, this project, in order to better address the problem, is based solely on the prediction of homicides and attempted murders, crimes that tend to be reported with more consistency.

2 Literature Review

The application of machine learning models to predict criminal activity is not a novel idea in the field. In a similar work to this paper, [1] examined the performance of K Nearest Neighbor (KNN) and Decision Tree algorithms to help predict crime in the city of Vancouver. However, their work was only able to refine a model that achieved roughly 40% accuracy, which is less than ideal for a problem as socially and politically sensitive as crime prediction. The problem that this analysis will handle will be a simpler variation that only deals with classifying homicide and non-homicide related cases so as to create a stronger and more useful model.

The concept of employing machine learning models to anticipate and prevent criminal activities has been further explored by [2], who proposed an integrated system that blends various surveillance techniques and machine learning methodologies, such as core analytics, neural networks, heuristic engines, recursion processors, Bayesian networks, data acquisition, cryptographic algorithms, among others. By adopting deep learning, machine learning, and computer vision

approaches, this system strives to create a more intelligent and efficient method for crime surveillance that emulates human thinking, yet operates non-stop and consistently. This advanced system could be utilized to predict crimes before they occur and to analyze crime scenes, possibly identifying elements that could be overlooked by human investigators. Moreover, the authors suggest incorporating scenario simulations, enabling the software to run multiple simulations of a given scenario based on 17 primary characteristics identified in their research, ultimately recommending an appropriate course of action or alerting police officials. The proposed integrated system demonstrates the potential of leveraging cutting-edge machine learning techniques in crime prediction and prevention, providing valuable insights for future research in this field.

Even today, there are substantial advances happening in the use of machine learning techniques in crime prevention. The further works in [3] involve the use of a Long Short Term Memory (LSTM) neural network, a far advanced technique in machine learning beyond what will be analyzed in this paper. The comparison between popular algorithms there show the rapid progression of neural networks in comparison to these simple classifiers, as LSTM was shown to outperform the other algorithms in prediction of crime at much finer nuances than will be analyzed in this paper. There is great interest and possibilities in the near future for breakthroughs in machine learning to help improve public safety.

3 Experimental Design

Since the data from the police reports divides multiple types of homicides into different categories, The signal_description column was recoded to compress all of the homicide categories into a single binary result, 1 representing a homicide police report and 0 if it is a non-homicide related report. This will allow us to predict whether a police report is a homicide or not given certain features.

As with a dataset like this, much of the data within each feature is NaN, which can pose a large problem when it comes to creating an accurate model. One may choose to impute the values, however this introduces a danger of contamination if our imputations are non-correlative. However, there are some features that could benefit from imputation. Since Victim and Offender Age are numeric values, it was fitting to simply impute the mean to all NaN values contained in the dataset. When it came to Victim Number and Offender Number, cases that had NaN meant that the crime involved 0 victims/offenders (or an unknown number), so an imputation of 0 was used for all missing values. For the rest of the categorical values, the data was too complex to impute a common "easy" value, so UNKNOWN was imputed in place of NaN. With this method, we are able to use much more of the data than before with our models, which will substantially increase accuracy.

For all categorical features, the data was one-hot encoded so as to allow for a model to digest the information properly.

The only features unused were ID related features such as ItemNumber and OffenderID. While Location and Time would be useful features, their one-hot encoding were found to be too large to be used in our models, so they were dropped.

To reduce the dimensions of our dataset to a more manageable size, we used Random Forest to select the most important features. A cutoff of 130 features was set to retain at least 99.5% of the original dataset's variance. The graph in Figure 1 shows various feature cutoffs that allowed us to come to the 130 feature cutoff conclusion that reduced the dimensionality of the data as much as possible while preserving the variance.

The analysis will involve the usual training and test split that is popular in the field of machine learning, but we will also use cross-validation to further evaluate the generalization.

The result of the pre-process was a dataset with 93,008 instances and 130 features.

4 Algorithm Selection

4.1 Logistic Regression

Logistic Regression is a popular algorithm for binary classification tasks, where the goal is to predict one of two possible outcomes. It is based on the concept of using the logistic function to model the probability of an event occurring. The logistic function, also known as the sigmoid function, transforms input values into a probability range between 0 and 1, which can be interpreted as the likelihood of an event happening. Logistic Regression assumes that the relationship between the log-odds of the response variable and the predictor variables is linear. It is chosen for this analysis because of its simplicity, ease of interpretability, and efficiency in handling small to moderately sized datasets. Additionally, it allows for a straightforward examination of feature importance through the coefficients of the model.

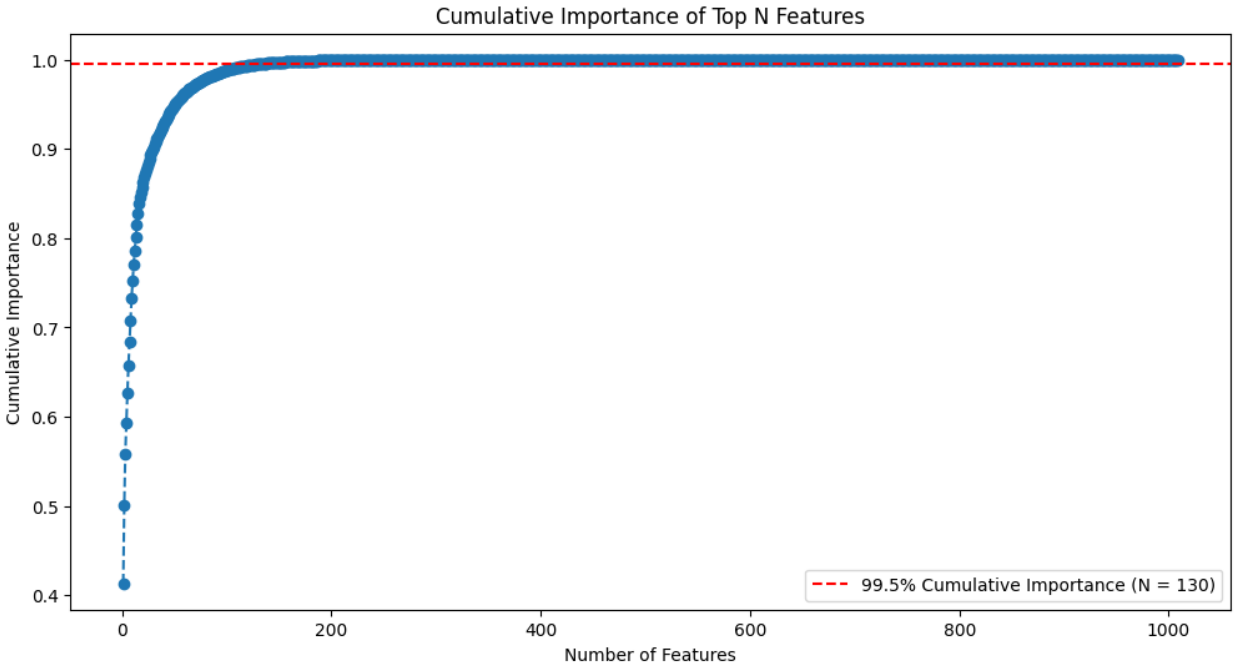

Figure 1: Data Variance by Feature Cutoffs

4.2 Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve the overall accuracy and robustness of predictions. Each tree in the forest is built independently, and the final prediction is based on the majority vote of all trees. Random Forest operates under the assumption that a group of weak learners, in this case, decision trees, can be combined to form a strong learner that is more accurate and stable. The algorithm is chosen for this analysis due to its ability to handle high-dimensional data, provide a natural mechanism for feature selection, and reduce the risk of overfitting. Furthermore, it can capture complex interactions between features, which can be particularly useful in understanding the intricacies of crime data.

4.3 SVM

Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification and regression tasks. The main idea behind SVM is to find the optimal hyperplane that separates the data points belonging to different classes with the maximum margin. The margin is the distance between the hyperplane and the nearest data points from both classes, known as support vectors. SVM makes use of kernel functions, such as linear, polynomial, and radial basis function (RBF) kernels, to transform the input data into a higher-dimensional space, enabling the algorithm to capture complex patterns and relationships between features. SVM is chosen for this analysis because of its ability to deal with high-dimensional and non-linear data, its robustness against overfitting, and its effectiveness in handling both small and large datasets.

5 Evaluation and Results

Model	Accuracy	Precision	Recall	F1 Macro (Avg CV F1 Score)
Logistic Regression	1.00	1.00	1.00	0.9983
Random Forest	1.00	0.99	0.88	0.9644
SVM	1.00	1.00	1.00	0.9974

Table 1: Performance Metrics for Each Model

The results of the three models indicate that both the Logistic Regression and Support Vector Machine (SVM) models achieved perfect performance on the test dataset, while the Random Forest model showed slightly lower performance.

Figure 2: Chart of Performance Metrics of Each Model

In the case of Logistic Regression, the model yielded an accuracy of 1.00, with perfect precision, recall, and F1-score for both classes. The cross-validated F1 Macro score was 0.9983, which supports the strength of this model. The excellent performance of Logistic Regression suggests that the relationship between the features and the target variable in the dataset can be well-represented by a linear model.

The Random Forest model, on the other hand, exhibited slightly lower performance, with an accuracy of 1.00, but an F1-score of 0.93 for the homicides class. Despite the relatively lower performance in comparison to Logistic Regression and SVM, the model still achieved high precision and recall values. The cross-validated F1 Macro score for Random Forest was 0.9644, indicating that the model's performance is consistent across different data splits. The difference in performance between the Random Forest and the other two models may be attributed to the presence of a strong linear relationship between features and the target variable, which the tree-based model might not capture as efficiently as the linear models.

The SVM model achieved perfect performance similar to the Logistic Regression model. The accuracy, precision, recall, and F1-score for both classes were all 1.00, indicating that the SVM model can perfectly discriminate between the two classes. The cross-validated F1 Macro score was 0.9974, extremely close to Logistic Regression's cross-validated F1 Macro score, which further reinforces this model's exceptional performance. The success of the SVM model may be attributed to its ability to find the optimal decision boundary that separates the two classes in the dataset, especially given that the hyperparameters used were generally meant for a linear dataset.

The graph shown in Figure 3 is described in this paragraph. The top features and their respective weights in each model depend on the learning approach used by the algorithm. While all three models highlight Signal_Type_30S as a crucial

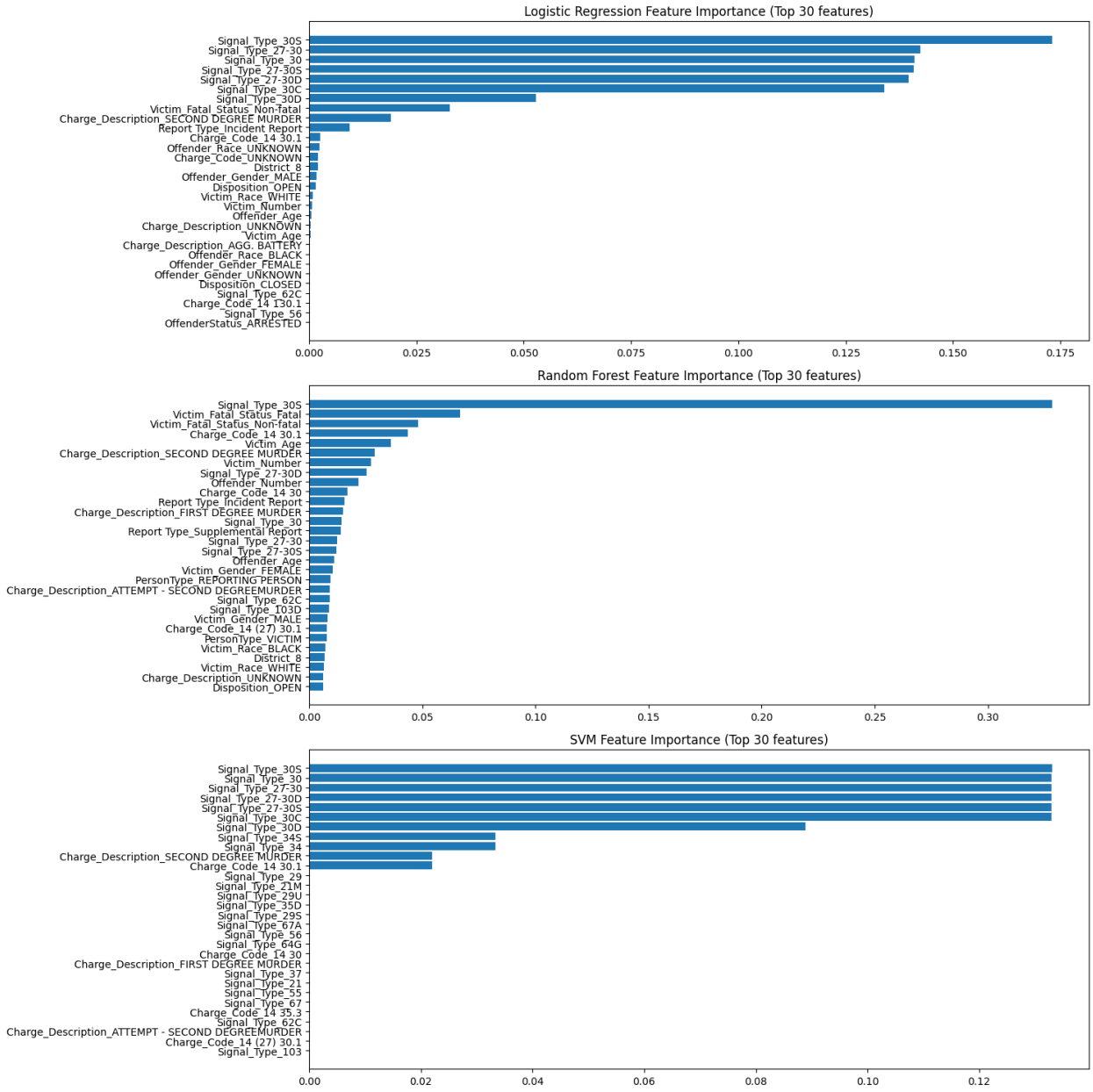

Figure 3: Top 30 Most Important Features of Each Model

feature, the other important features vary based on how the model learns from the data and how the features contribute to the model's decision-making process. An interesting thing to note here is that Random Forest appeared to apply at least some importance to many more features than the other two more linear based models, which may have lead to its downfall in performance due to overfitting issues. Logistic Regression's L1 regularization can be seen here as well with the stark dropoff in feature weights beyond the first 10 or so features. However, overall the models generally had similar weight orderings of each feature, suggesting there are generally strongly correlative features toward the classifications in this dataset.

The high performance of all three models demonstrates that the chosen features are highly predictive of the target variable. The perfect scores achieved by the Logistic Regression and SVM models indicate that a linear decision boundary is sufficient to separate the two classes, while the slightly lower performance of the Random Forest model suggests that the tree-based model might not be as effective in capturing the linear relationship between the features and the target variable due to potentially noisy features.

6 Conclusion

This report has provided valuable insights into the significance of various features in predicting crime-related outcomes using different machine learning models. Despite the hypothesized non-linear result of the dataset, we found that it was much more linear than expected, suggesting strong predictive capabilities of these models in other crime environments. With proper tuning, we found that all three models were extremely suitable for this task so long as the data is pre-processed well. By understanding the strengths and limitations of each model and the underlying relationships between features and the target variable, we can develop more accurate and robust predictive models for crime prevention and related applications.

Future works could serve to explore the usage of more complex models. As mentioned in the literature review, neural networks have untapped potential in understanding very nuanced differences that appear in crime data, which begs further research to create a more refined model that could apply to a wider variety of cases. With a more powerful and well-trained model, predictions could be made on a day or even hourly granularity, greatly increasing the capabilities of law enforcement everywhere.

The real-life implications of this analysis could be substantial, particularly in the area of crime prevention and law enforcement. By identifying the most influential features in predicting crime outcomes, policymakers and law enforcement agencies can prioritize resources and develop targeted interventions to address the root causes of crime. For instance, understanding the significance of specific signal types or victim demographics can help tailor community-based programs or allocate law enforcement resources more effectively. Moreover, the development of accurate predictive models can facilitate proactive policing strategies and enable law enforcement to identify potential crime hotspots or at-risk individuals, allowing for early interventions and ultimately leading to a safer society.

References

[1] Suhong Kim, Param Joshi, Parminder Singh Kalsi, and Pooya Taheri. Crime analysis through machine learning. In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 415–420, 2018.

[2] Neil Shah, Nandish Bhagat, and Manan Shah. Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. *Visual Computing for Industry, Biomedicine, and Art*, 4(1), April 2021.

[3] Xu Zhang, Lin Liu, Luzi Xiao, and Jiakai Ji. Comparison of machine learning algorithms for predicting crime hotspots. *IEEE Access*, 8:181302–181310, 2020.