Robust Modelling of Crime Prediction

Danial Kamali, Sachit Gaudi

April 29, 2024

1 Introduction

The research project aims to enhance law enforcement intelligence to better equip police for crime prevention. The primary emphasis lies in predicting crimes, a crucial component of law enforcement intelligence. However, current techniques are very limited to the frequentist approaches, including identifying crime hot spots [3]. The major challenge to predictive modeling is the diverse and multi-modal nature of the data, encompassing categorical, continuous, temporal, and geospatial features. This study aims to address this challenge by comprehensively evaluating classical machine learning algorithms across this broad range of data types. The ultimate goal is to establish a robust benchmark that can be used to critically assess the effectiveness of various machine learning algorithms for crime prediction.

Crime is a pervasive issue in our society that has an impact on everyone, whether as a victim or perpetrator. In our project, we examined the "Chicago Crime Dataset" which contains information on criminal incidents in Chicago dating back to 2001. Our analysis focused on identifying patterns and trends in crime over time, as well as pinpointing the police districts where crime occurs most frequently to identify the hotspots of some specific criminal activity.

The main objective of this project is to build a predictive model that can predict the type of crime based on preliminary information received from the reporter such as location, time, and description of location. The code is available here [1]. Through this project, we aim to learn about the following:

[1] https://github.com/iamdanialkamali/Crime-Prediction/

1. How do we incorporate features of different modalities into classical machine learning models?
2. Examination of the Significance of Features in Predicting Crime Types within the Context of Chicago
3. What are engineering modifications we need to make to handle large scale data?
4. Strategic approaches for addressing temporal distribution shifts within the dataset.
5. Evaluation of methodologies aimed at mitigating class imbalance concerns, with a focus on achieving equitable error rates across diverse classes.
6. Which classification algorithm performs best for this prediction task?
7. How accurately can we predict the type of crime from preliminary information in Chicago?
8. Can we identify any patterns or trends in the data that can help us prevent and respond to crimes more effectively?

By answering these questions, we can gain insights into the nature of crime in Chicago and help law enforcement officials make data-driven decisions to prevent and respond to crimes.

2 Problem Statement

The Chicago crime dataset contains information on various reported crimes in the city of Chicago from 2001 to the present lastly updated on October 26th, 2023. This project aims to predict the type of crime based on location and time. By accurately predicting the type of crime, law enforcement officials can take appropriate measures to prevent and respond to crimes in a timely and effective manner.

The inherent characteristics of crime data remain largely unexplored. Our initial research strategy is a thorough examination of the data, as outlined in Section 4.1. In this section, we emphasize the significance of conducting

data exploration, a fundamental step in gaining insights into the data distribution and identifying potential discrepancies in relation to the underlying assumptions. Subsequently, we comprehensively analyze these assumptions in Section 4.2. These nature of assumptions will lead us to find new techniques to leverage challenges in data such as Distribution shift and class imbalance. The detailed approach has been listed in 4.4 and 4.6

Furthermore, our investigation extends to studying diverse machine learning algorithms, each applied within specific contextual conditions, as elaborated in Section 4.7.

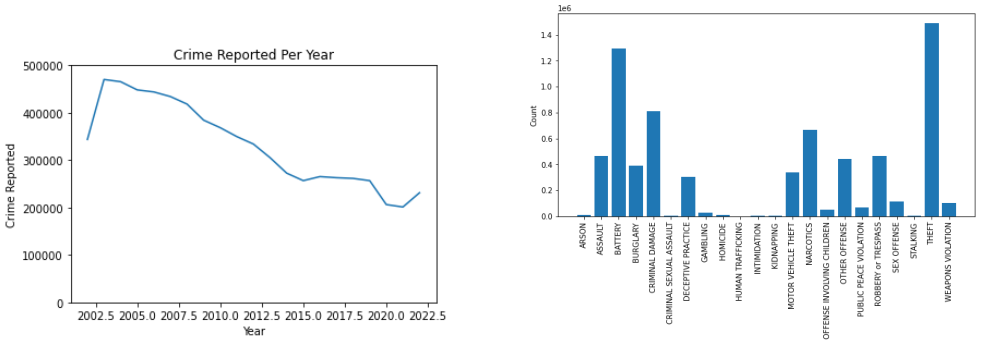
The success of deep learning can be attributed to learning a map of multi-modal data into tailored functional spaces, which enhances its representational capabilities. Conversely, feature space is assumed to be infinite-dimensional Hilbert space and is approximated by the kernels in classical machine learning. Our research delves into methodologies for data representation and feature engineering, as detailed in Section 4.3. While augmenting features can indeed bolster a model's representational power, it comes with a caveat known as the "curse of dimensionality," which can hinder a model's ability to generalize effectively. In Section 4.8, we explore several strategies, including feature selection and regularization, and discuss their potential to enhance model robustness, noting the connection between regularization and selection.

In Section 5, we establish the experimental framework, a critical component for evaluating model performance. In testing our algorithms, we employ a robust test set that involves a temporal separation between training and testing data. This setup assesses the model's proficiency in predicting future occurrences of crimes. However, it is essential to acknowledge that this stringent testing criteria necessitates an assumption that there is no shift in distribution. Addressing distribution shifts and their integration into the algorithm can be challenging, and while we acknowledge this issue, we aim to limit our commentary on distribution shifts given the extensive nature of this topic.

To accomplish this objective, we have subdivided the task into the following subtasks, where we aim to explore classical machine learning algorithms and propose enhancements to develop a robust predictive model. It's important to note that this approach is not constrained solely to crime data but can be applied to a wide range of datasets.

3 Dataset

The Chicago crime dataset [2] contains information on various reported crimes in the city of Chicago from 2001 to the present last updated on October 2023. The Chicago Crimes - 2001 to present dataset provided by the City of Chicago Police Department is available on the City of Chicago's data portal[2] and will be used for this project. This dataset contains about 7.76M lines of information on reported crimes in the city, including the type of crime, location of the crime, date and time of the crime, etc [Table 4]. As of October 2023, this dataset has over 7.5 million records and 22 columns. There are, on average, 400,000 reported crimes [Figure 1] for a population of 2.697 million, which is a big and steady problem to solve.



(a) How crime progressed throughout the years

(b) Nature of crime

Figure 1: Description of Crime Dataset

4 Approach

4.1 Looking into data

The analysis of crime patterns and trends in urban areas has important implications for law enforcement, public policy, and social welfare. In the case of Chicago, crime is a persistent problem that affects the safety and well-being of its citizens, as well as the reputation and economic vitality of the city. Therefore, the study of crime in Chicago is of great interest to researchers, policymakers, and the public.

[2]https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

One of the primary objectives of our analysis was to examine the trends in crime rates over time. By analyzing the Chicago Crime dataset, we found that the overall crime rate has decreased over the years, as shown in Figure 1. This trend is an encouraging sign that crime prevention efforts and law enforcement strategies have been effective in reducing crime in the city. However, the sudden drop in crime during the COVID-19 pandemic and its subsequent resurgence after the end of quarantines is a point of concern that merits further investigation.

Moreover, we also identified the hot zones of crime in Chicago by analyzing the distribution of crime across the 24 police districts, as shown in Figure 4. This information can help law enforcement agencies to allocate resources and personnel to areas that are more prone to crime and to develop targeted crime prevention strategies. For instance, the police department can prioritize specific crimes or subjects to focus on in each district or assign officers to districts based on their specialty or expertise.

In addition to identifying the high-crime areas of the city, we also analyzed the nature of crimes in each district, as shown in Figure 5. The distribution of crimes in each district varies significantly, which can provide valuable insights to law enforcement agencies in developing strategies to combat crime. For example, they can identify the types of crimes that are most prevalent in a particular area and prioritize their efforts accordingly. Moreover, they can provide additional training to police precincts in high-crime areas to improve their ability to respond to and prevent crime.

Another important aspect of our analysis was to examine the patterns of crime across different months of the year. We hypothesized that environmental factors such as weather might influence the nature and frequency of crimes. However, our observations did not support this hypothesis, as shown in Figure 6. The distribution of crimes remained relatively consistent throughout the year, indicating that other factors such as socioeconomic conditions or social norms may play a more significant role in shaping crime patterns.

Finally, we analyzed the distribution of crimes across different times of the day, dividing the 24-hour day into four phases. As expected, the overall crime rate is lower at night, when most people are sleeping. However, we found that the distribution of crimes varies significantly across different parts of the day, as shown in Figure 7. For instance, battery crimes are more prevalent late at night, while robberies are more likely to occur in the early evening. This information can help law enforcement agencies to allocate resources and

personnel to areas and times of day that are more prone to crime, and to develop targeted crime prevention strategies.

4.1.1 Data Pre-Processing

From the information available in the Chicago Crime dataset, lots of them are unrelated to our goal such as the arrest information and IUCR. In addition, we have duplicate information such as zip code and their Latitude or beat information. Hence we removed this information from our dataset {'ID',' Case Number', 'IUCR', 'Arrest', 'Longitude', 'Domestic', 'Beat', 'FBI Code' , 'Updated On', 'Latitude', 'Historical Wards 2003-2015', 'Zip Codes', 'Location'}. These columns were either unrelated or had other columns representing them. Since it is a real-world dataset it has missing entries (NA). We removed these rows which were less than 1% of our dataset. Furthermore, most of our data were string data types. We numerically classified them by an integer index as their new representation.

4.1.2 Feature Engineering

We create new quantified features that can provide additional information about the crime occurrence. We extracted the information about the day of the week, month, and time of day from the Date column. This could help us extract more meaningful information from the dataset and improve the accuracy of the model.

4.1.3 Feature Selection

In our system, we assumed that we can extract the exact location of the crime and extract all other information about the location such as neighborhood, community and etc. Thereby, we start with these 19 features.{'Y Coordinate', 'Weekday', 'Description', 'Police Beats', 'Zip Codes', 'Census Tracts', 'Location Description', 'Month', 'Primary Type', 'X Coordinate', 'Wards', 'Time of Day', 'Community Area', 'Police Districts', 'Ward', 'District', 'Year', 'Block', 'Community Areas'}. We used techniques such as Sequential Feature Selection and correlation analysis to select the most relevant features for the prediction task. This will help us reduce the dimensionality of the data and improve the performance of the model. these methods discarded features like the 'Month' feature from our feature set. Figure 6 shows the histogram distribution of the crime types in different months of the year.

As we can see in Figure 6, the distribution of crimes stays similar around every month of the year. Here are the features selected after the feature selection. {'Year', 'Description', 'Y Coordinate', 'Block', 'Ward', 'Location Description', 'Census Tracts', 'X Coordinate'}

We also experiment with using regularizer to select features, and SVD is also one of the important technique to redce the diamentionality and perform feature selection in a different subspace and project back to the original subspace.

Method	**Train Acc (%)**	**Test Acc (%)**	**Generalization Gap (%) ↓**
None	99.98	93.15	**6.83**
SFS	99.55	92.30	7.25
Hinge Selection	99.98	93.15	6.83
SVD (0.9 variance)	99.98	87.87	12.11

Table 1: Performance Metrics for Different Methods for feature selection applied for Random forests Classifier

The table reveals that none of the feature selection approaches proved effective, likely due to the dataset's substantial size relative to the number of features. Additionally, the application of SVD further deteriorated performance, likely attributed to the low variance in the timing of certain crimes, with a majority occurring during the night.

This low variance in the timing of crimes poses a challenge, particularly in cases where the removal of the time variable results in a loss of predictive power. For instance, omitting the time of the crime could compromise the model's ability to distinguish between day crimes, such as bank robberies and other minor offenses, versus serious crimes.

4.2 Assumptions

Geospatial data, characterized by latitude and longitude coordinates, deviates from the assumptions of Euclidean space. Effective handling of such data may necessitate either a transformation of the coordinates or the utilization of algorithms specifically tailored for spherical geometry. In our approach, we propose the conversion of geospatial data into Euclidean coordinates.

The logistic regression model commonly assumes a uniform error distribution. However, this assumption is not consistently met in our dataset,

primarily due to the presence of categorical variables. Furthermore, the lack of ordering in these categorical variables poses a unique challenge. If the categories were organized as $1, 2, 3 \cdots$, we could not infer a Euclidean meaning from the categories. Traditional one-hot encoding is not a viable solution due to the large number of unique values in the categorical feature, resulting in a data explosion of up to 1 TB. Additionally, the challenges of overfitting associated with a large number of features further complicate matters.

To address these issues, we propose the use of feature embeddings, as expounded in Section 4.3. These embeddings will project categorical variables into a high-dimensional space, thereby maximizing predictive power and improving our ability to discern the nature of criminal activities. Notably, our adoption of embeddings has yielded a substantial 32% increase in model accuracy compared to using traditional categorical variables, as embeddings obviate the need to consider the order among categories.

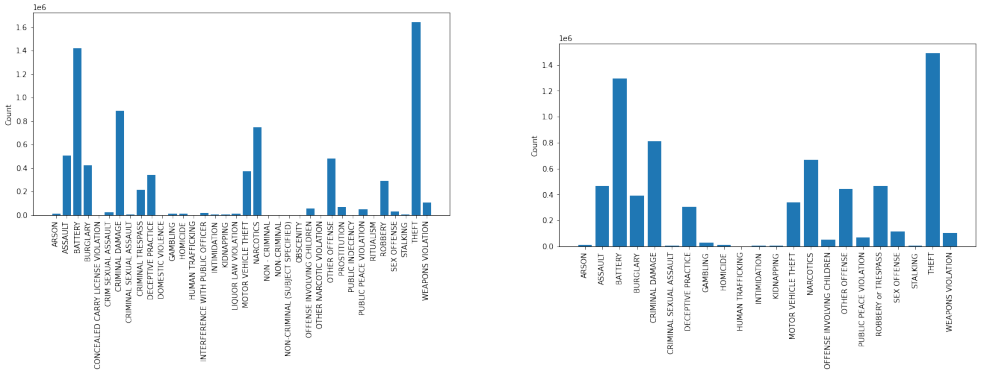
4.3 Multi-modal Data

The transformation of text, location, and categorical variables into Euclidean space is imperative for our analysis. To address this complex task, we propose the utilization of embeddings. Specifically, for text features, which often exhibit substantial dimensionality, we will employ BERT embeddings. To manage the high dimensionality of these text features, we plan to perform dimensionality reduction to achieve the desired representational power.

For categorical variables, we aim to train embeddings that maximize their representation of the predicting variable (nature of crime). This will be accomplished by training a linear Multi-Layer Perceptron (MLP) on top of the embeddings, utilizing cross-entropy loss with respect to the predicting variable. This approach ensures that the embeddings capture the intricate relationships between categorical variables and the target variable, enhancing their predictive capabilities.

These classifiers trained with embeddings instead of categorical variables yield a substantial increase in accuracy, amounting to an impressive 32%.

4.4 Imbalance: Over, Under, and Generative sampling



(a) Distribution Crime Types (b) Normalised Nature of crime

Figure 2: Distribution Crime Types After Label Normalization

Figure 2 illustrates the highly imbalanced class distribution of the Chicago Crime dataset labels. In order to address this issue, we performed some data preprocessing steps, including the removal of non-criminal labels and the merging of similar labels, as shown in Table 2. The new disribution is shown in Figure 2

New Label	Old Labels
NARCOTICS	OTHER NARCOTIC VIOLATION NARCOTICS
SEX OFFENSE	PROSTITUTION CRIM SEXUAL ASSAULT SEX OFFENSE
GAMBLING	RITUALISM LIQUOR LAW VIOLATION GAMBLING
ROBBERY or TRESPASS	CRIMINAL TRESPASS ROBBERY
PUBLIC PEACE VIOLATION	INTERFERENCE WITH PUBLIC OFFICER PUBLIC PEACE VIOLATION

Table 2: Label Conversion [6]

Following the modification of class labels, a substantial class imbalance persists, as evident from the low macro F1 score and a high accuracy. Addressing this challenge and aligning the F1 score more closely with accuracy necessitates the exploration of various sampling techniques.

Among these techniques, oversampling is a popular choice. However, in scenarios where the dataset is already extensive, employing oversampling alone could result in a 40-fold increase in data volume. This amplification not only leads to slower training times but can also be impractical for certain algorithms, such as Kernel SVM, which may encounter prolonged execution times. Moreover, oversampling in isolation may foster overfitting for the minority class. To mitigate these issues, it is recommended to combine oversampling with other techniques, such as cross-validation. Cross-validation involves dividing the dataset into training and testing sets, enabling model evaluation on the testing set to prevent overfitting and ensure robust generalization to new and unseen data. Selective oversampling, as illustrated in our example, contributes to notable improvements in the F1 score.

Another technique, undersampling, is employed to manage overfitting and reduce data size. However, this approach becomes problematic when dealing with classes that have very few samples, as is the case with serious crimes occurring infrequently. The resulting dataset becomes highly sparse, diminishing generalizability and rendering it unusable.

To overcome both the challenges of data multiplicity and maintaining representational capacity, we explore generative sampling. This technique assumes that the features of each class follow a specific distribution (e.g., Gaussian in our example) and that each feature is independent of others given the class. While these assumptions can be stringent, our example highlights that generative sampling did not yield superior results, primarily due to the invalidation of the independence assumption among features. Fine-tuning these assumptions is crucial for achieving optimal results in such scenarios.

4.5 Large Scale Dataset

The dataset, consisting of 7.7 million data points amounting to 2 GB, poses a formidable challenge for the application of classical machine learning algorithms such as SVM and random forest. To overcome this engineering obstacle, we propose two viable solutions.

Firstly, we suggest leveraging approximate methods for computing distance/similarity, utilizing the faiss library developed by Meta, specifically

designed to handle large-scale data.

The second strategy involves temporal data chunking—feeding smaller subsets to the classifier and ensemble these classifiers using boosting techniques. This approach offers dual benefits: it proportionally reduces training time, facilitating parallelization and significantly decreasing the overall training duration. Additionally, it introduces an automatic re-weighting mechanism, where the ensemble classifier comprises all individual classifiers trained on a single data chunk. The weight of each classifier is inversely proportional to its logarithmic error rates, akin to boosting. This automatic re-weighting mechanism proves advantageous for addressing distribution shift, particularly when the validation set is temporally aligned with the test set. The validation set aids in re-weighting samples to better align with the distribution of the test set, thereby enhancing overall accuracy.

Notably, the chunking technique has resulted in a substantial 6% improvement in test accuracy, effectively bridging the gap between test and train accuracy and serving as an effective regularizer.

4.6 Distribution Shift

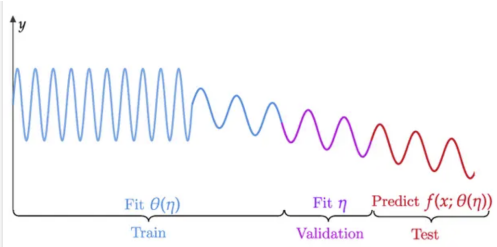


Figure 3: Shift in distribution

As evident from Figure 1, both the crime rate and the nature of criminal activities exhibit dynamic changes from year to year. Constructing a robust model based solely on historical data poses challenges in adapting to future distributions. To address this issue, we recognize the necessity of accounting for distribution shifts in our approach.

One strategy involves assigning differential weights to data points based on their temporal proximity to the future. Alternatively, we propose employing the chunking technique, as detailed in Section 4.5. This technique eliminates the need for manual weight assignment and tuning by automatically

assigning weights according to the significance of data points in predicting outcomes on the validation set, which is selected to closely resemble the test set.

In particular, we advocate assigning weights inversely proportional to the logarithm of error rates, akin to the principles of boosting.

$$\alpha_m = \frac{1}{2} \log \frac{1 - e_m}{e_m}$$

This method aims to dynamically adapt to the evolving nature of the data distribution and enhance the model's resilience to temporal shifts.

4.7 Comparison of Classification Algorithms

It's important to choose the right machine-learning algorithm that best fits the nature of your data and the problem you're trying to solve. Logistic regression may not always be the best choice, especially if the data violates its assumptions. The data is prone to human error and false reporting of the crime, So we need to study the quality of the data; if we have more outliers, then a model like SVM, which is robust to outliers, might be a better fit. We will test different classification algorithms such as decision trees, random forests, and support vector machines to predict the type of crime based on location and time. This will help us identify the best algorithm for this prediction task. We will compare the performance of the different classifiers using appropriate metrics such as accuracy, precision, recall, and F1-score. This will help us identify the best model for predicting the type of crime based on location and time.

4.8 Generalisation, Regularisation & Selection

We cannot increase features as discussed above, We need to select a set of features that impact the most. The general technique to explore feature selection is consider all the possible combination of features, which will be of the order 2^K, which is not feasible to achieve. Forward selection, Backward selection, mutual information and correlation analysis algorithms will achieve polynomial order $O(K)$ by using heuristics. but each of the algorithms are known to have limitations.

Regularisation is one of the well studied aspect of machine learning to improve generalisation. We will study the feature relevancy using regular-

isation. L1 regularisation is known to nullify some features, for a given hyper-parameter resulting in feature selection. We are interested to find an answer for whether regularisation provide better feature selection compared to the classical methods in $O(K)$ or $O(1)$.

We will try to leverage the auto regularisation and the robust to outlier nature of SVM. However, SVM suffers from its own problems. The data imbalance Figure 1 will make SVM unstable as studied by [1]. Kernel methods needs huge memory in order of N^2, where N being the number of data samples. given the samples in millions it is practically impossible to compute kernel based functions. We need to further study sampling based tricks to effectively use SVM. This makes a statement that generalisation is not a free lunch.

.

5 Experimental Setup

5.1 Cross-Validation

In case we face any accuracy issues, we will evaluate the performance of different classifiers using cross-validation techniques. This will help us estimate the accuracy of the model and detect any over fitting or under fitting issues.

5.2 Hyper-parameter Tuning

We will optimize the hyper-parameters of the chosen algorithm to achieve the best performance. This will help us fine-tune the model and improve its accuracy.

5.3 Metric

In addition, we need to use a metric that, in nature, is more sensitive to the underrepresented classes. Hence, we use the macro-averaged F1 (Macro-F1) metric for our evaluation.

6 Results

We test three different classification algorithms such as decision trees, random forests, K-NN, Logistic Regression, MLP to predict the type of crime. This will help us identify the best algorithm for this prediction task. We will compare the performance of the classifiers using the f1 score metric. We will use 5-fold cross-validation to help us identify the best model.

For each classifier introduced, we incorporate our proposed methods and conduct an ablation study to comprehensively assess how our approaches, as outlined in Section 4, align with each classifier. Subsequently, we delve into a detailed discussion regarding the implications of these approaches, as expounded in Section 7.

Model	Transformation	Train Accuracy	Test Accuracy	F1-score	Time
LogisticRegression	-	11.59	9.03	0.75	616.34
	L1	19.16	18.56	0.91	583.45
	L2	19.21	18.51	0.99	535.10
SVM	Linear kernel	19.21	18.51	0.99	1571.28
	RBF	24.46	22.32	1.21	2500.11
KNN	faisis	-	53.43	28.72	0.12
Decision Trees	Random Forest	99.98	93.15	88.56	81.58
	Random Forest (No Embedding)	79.99	60.27	43.26	64.79
	Random Forest + Over Sampling	99.99	92.97	91.26	43.01
	Random Forest + Under Sampling	100	69.36	49.44	0.91
	Random Forest + Generative Sampling	100	3.88	2.74	52.21
	Random Forest + Chunk	99.99	91.72	73.89	**2.24**
	Random Forest + Generative + Chunk	100	3.81	2.68	183.78
	Random Forest + Distribution Shift + Chunk	99.98	91.91	72.86	22.47
	Gradient Boosting	88.12	88.66	88.02	2224.26
	Light Boosting [LGBMClassifier(learning_rate=0.01, num_leaves=100)]	93.92	93.86	83.38	31.01
	Chunk + Light Boosting	93.86	92.85	82.04	27.01
	Distribution Shift + Light Boosting	93.83	**93.61**	**92.77**	31.21
MLP	-	22.10	21.61	1.61	262.92
	Chunk + L2	21.61	20.81	1.56	81.6
	Distribution Shift + Chunk + L2	21.60	20.81	1.56	77.4

Table 3: Results

The heatmap in Figure 8 displays the confusion matrix of our best-performing classifier. From the heatmap, it is evident that most of the misclassifications occur in the STALKING, ASSAULT, or BATTERY categories. These categories are semantically similar, which could make it difficult

for the model to distinguish between them without additional information. Additionally, we observe that the oversampling method only improves the performance of the RandomForest classifier. This could be due to the highly imbalanced labels, which may render the oversampling method ineffective for other classifiers.

7 Summary and Conclusions

We first introduced the Chicago crime dataset and described its importance as a valuable source of information on crime patterns in the city. We then provided statistics about the data and experimented with our hypotheses on the data. Our analysis focused on various aspects of the dataset, including the trend of crime rate over the years, the distribution of total crime in each police district, and the nature of crimes in each district. We also investigated whether the crime statistics changed according to the month of the year and the time of the day.

In the subsequent section on data pre-processing methods, encompassing dimensionality reduction and feature selection, our findings align with the analyses presented in the preceding section. We demonstrate that the implemented pre-processing techniques contribute to enhanced classifier performance.

Subsequently, we delve into the challenges inherent in the dataset, encompassing issues such as class imbalance, distribution shift, multimodality, and scale. To address these challenges, we propose a range of approaches aimed at mitigating the identified issues and improving the robustness of our models.

In the concluding phase, we conducted a comparative analysis of the performance of four distinct classifiers—Random Forest, Logistic Regression, SVM, and Decision Trees—specifically applied to the task of crime prediction. Through an ablation study, we aimed to discern the optimal combinations of these approaches under varying settings. The following key insights summarize our findings:

- Decision Tree-based models perform well on large-scale classification.
- Gradient Boosting further brings the test performance close to training.
- Chunking helps to give more weight to recent data compared to older ones.

- Undersampling in large-scale data leads to catastrophic performance drop.
- Learned embeddings are more expressive compared to categorical features.
- In large datasets, regularization's importance diminishes; higher complexity boosts performance, and distribution shift plays a pivotal role.

Overall, our work contributes to the growing body of literature on crime prediction and data analysis by providing insights into the Chicago crime dataset. We believe that our findings can be used to develop effective strategies for crime prevention and to inform policy decisions aimed at reducing crime rates in the city of Chicago.

Among the notably successful algorithms, tree-based methods exhibit promise, mirroring the decision-making processes employed by law enforcement officers. While individual trees are characterized by relatively low variance, bagging and boosting techniques will enhance their overall performance.

For this work we used scikit-learn [5] library for machine learning algorithm, and seaborn [7] library for visualization.

8 Future Works

It is worth mentioning that our analysis of the Chicago Crime dataset focused mainly on examining time-related information. However, there are still many opportunities for future work with this dataset. One possible avenue for future research would be to integrate additional external data sources to enrich our understanding of crime patterns in the city. For example, demographic data like income, education, and race, or GIS models, could be used to identify patterns of crime in different neighborhoods and provide insights into the social and economic factors that contribute to crime.

In addition, more advanced machine learning techniques, such as deep learning and natural language processing, could be employed to analyze the unstructured data in the dataset, such as crime descriptions and location data. These advanced techniques can help to identify subtle patterns and relationships in the data that may not be immediately apparent using traditional machine learning methods like random forests, KNN, and ensemble methods. Future research could explore the use of these advanced techniques

to enhance our understanding of crime patterns and support the development of more effective crime prevention and intervention strategies.

The current state-of-the-art models for tabular data include TabTransformers, as introduced by Huang et al. [4]. Our embedding approach aligns with elements of this methodology, and we observe enhanced gains as we augment the model's capacity, leveraging the large scale of data in conjunction with relatively smaller features. Our interest lies in investigating how the model's capacity, specifically the number of layers, contributes to performance improvement.

References

[1] Rukshan Batuwita and Vasile Palade. *Class Imbalance Learning Methods for Support Vector Machines*, chapter 5, pages 83–99. John Wiley Sons, Ltd, 2013.

[2] Chicago (Ill.). Police Department. Chicago crimes, 2001-2018. Inter-university Consortium for Political and Social Research [distributor], 03 2019.

[3] Vikas Grover, Richard Adderley, and Max Bramer. *Review of Current Crime Prediction Techniques*, pages 233–237. 01 2007.

[4] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings, 2020.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[6] Vaibhav3M. Vaibhav3m/chicago-crime-analysis: Bigdata analytics on chicago crime dataset.

[7] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

9 Appendix

Feature	Type	Description
Location Description	Categorical	Describes the location where the crime occurred
X Coordinate	Geospatial	Longitude where the crime occurred
Y Coordinate	Geospatial	Latitude where the crime occurred
Community Area	Categorical	Area or neighborhood where the crime took place
Police Beats	Categorical	Police beat where the crime was reported
Wards	Categorical	Political division where the crime occurred
District	Categorical	Police district where the crime was reported
Year	Temporal	Year when the crime was reported
Date	Temporal	Exact date and time when the crime was reported
Description	Text	Detailed description of the crime
Primary Type	Categorical	The main category of the crime
Census Tracts	Categorical	Census tract where the crime occurred
Block	Categorical	Specific block or street where the crime occurred
Zip Codes	Categorical	Postal code of the crime location

Table 4: Snapshot of Chicago Crime dataset

Primary Type: Primary type of the crime committed. containing 36 prime types of crime. here is a small number of crimes types:

- ARSON
- ASSAULT
- BATTERY
- BURGLARY
- CONCEALED CARRY LICENSE VIOLATION
- CRIM SEXUAL ASSAULT
- CRIMINAL DAMAGE
- CRIMINAL TRESPASS

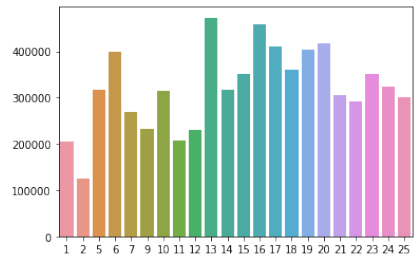


Figure 4: Crime per police districts

- DECEPTIVE PRACTICE
- GAMBLING
- HOMICIDE
- HUMAN TRAFFICKING

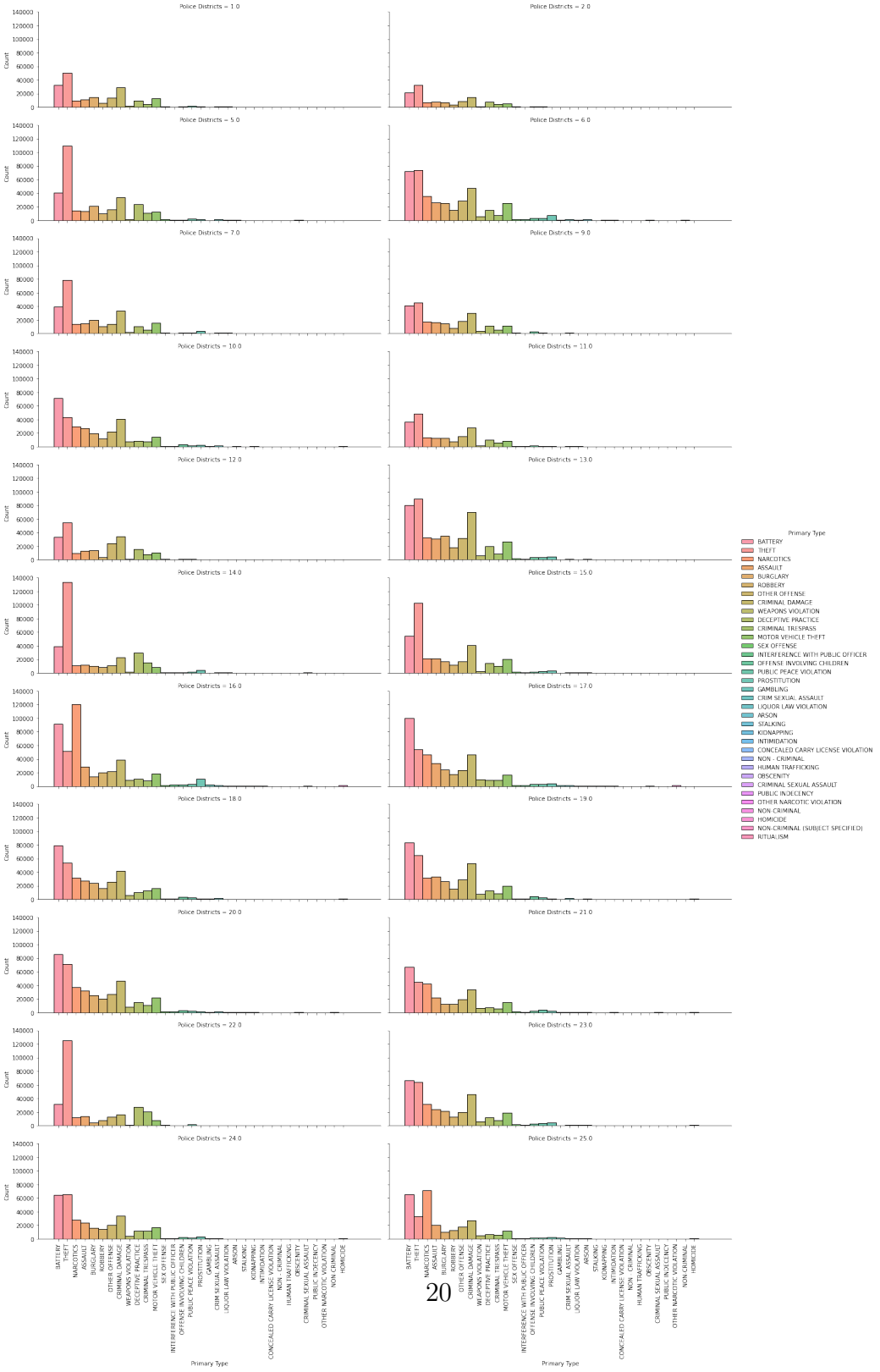


Figure 5: Distribution of crime per police districts

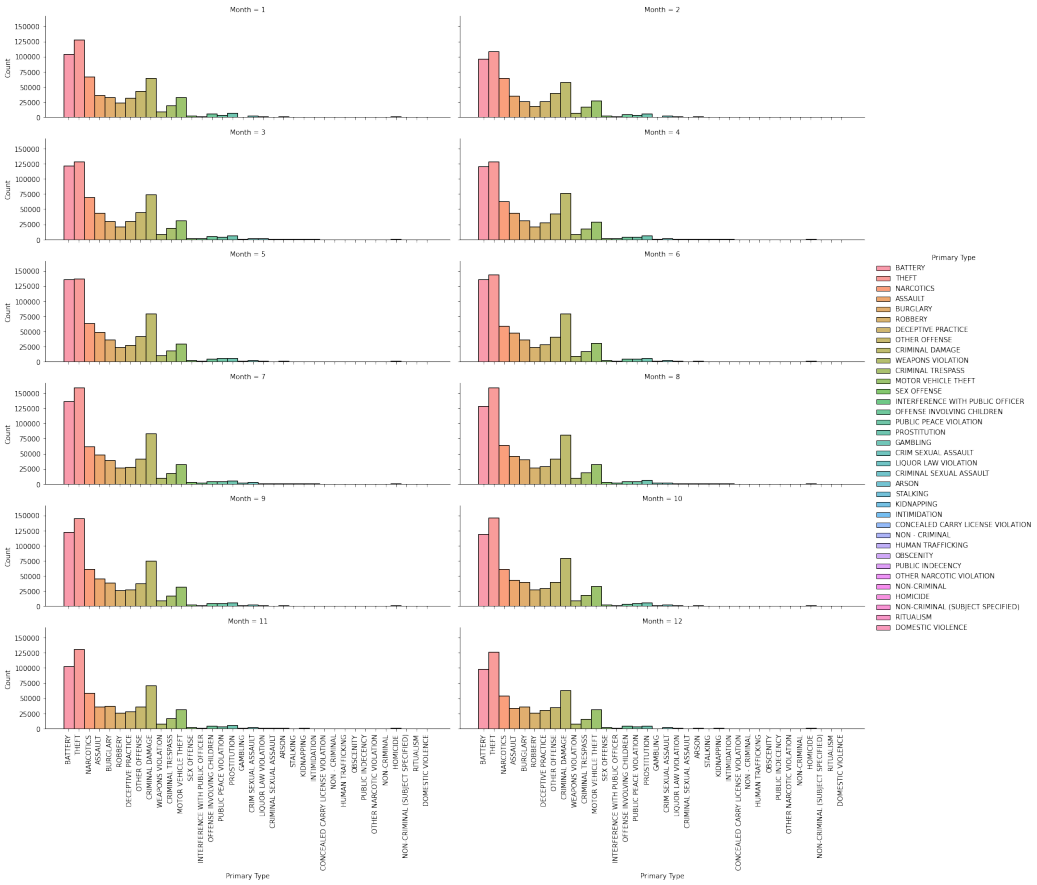


Figure 6: Distribution of crime per police month

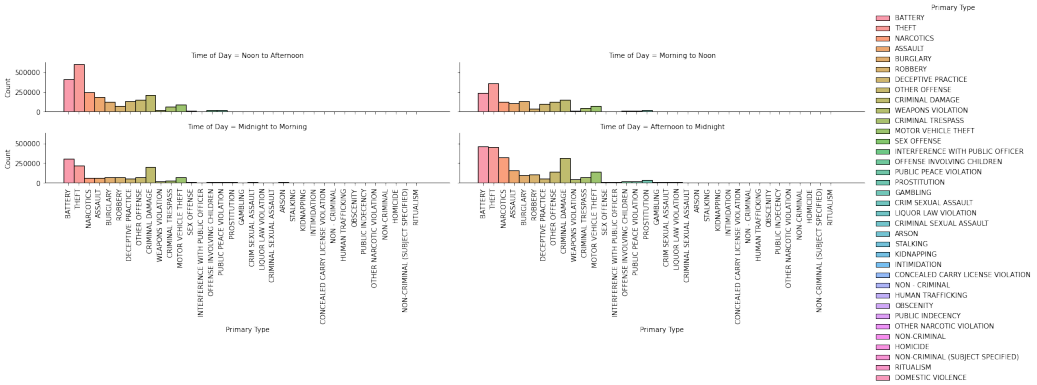


Figure 7: Distribution of crime at each time of day

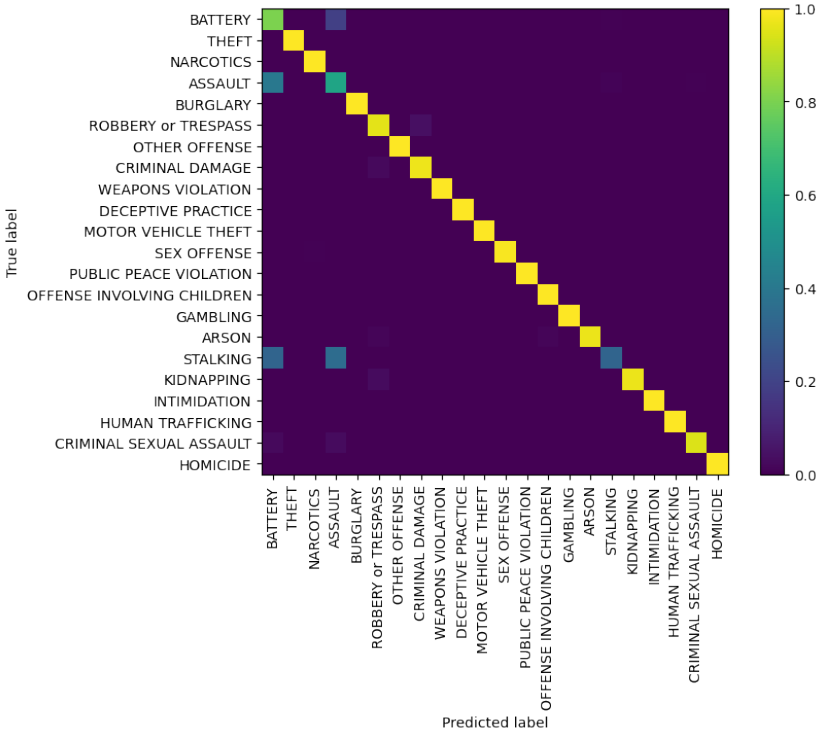


Figure 8: Normalized Confusion Matrix Heat map