IE 534 Final Project
Chicago Crime Prediction

Completed By -
Yash **K**alyani (yashk5@illinois.edu)
Yash **B**ajaj (yashpb2@illinois.edu)
Mayank **A**garwal (mayanka5@illinois.edu)
Debapratim **G**hosh (dg19@illinois.edu)

Agenda

1. Problem Description
2. Data Description
3. Exploratory Data Analysis
4. Challenges and Remedies
5. Data Preprocessing Steps
6. Model Architecture
7. Model Selection
8. Conclusions

Slide Contribution: Yash K, Yash B, Mayank A, Deb G

Problem Description

- According to USNews, violent crime rate in Chicago has been somewhat higher than the national average for the last 5 years, while property crimes have been consistently lower. Therefore, it has become necessary to accurately predict future crime in Chicago.
- This would enable an optimised allocation of resources for fighting crime and would ensure an overall reduction in crime rate of all types of crimes.
- This project aims at building a deep learning based multi-class classification model that can accurately predict the occurrence and type of crime in the future.

Slide Contribution: Yash K, Yash B, Mayank A, Deb G

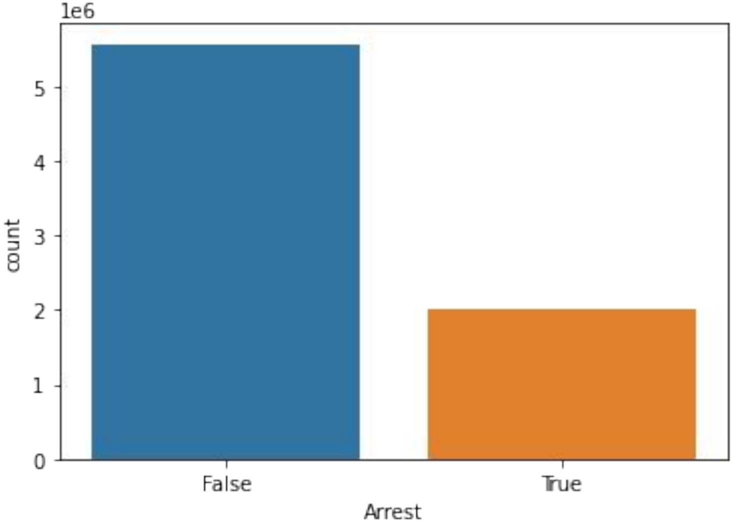
Data Description

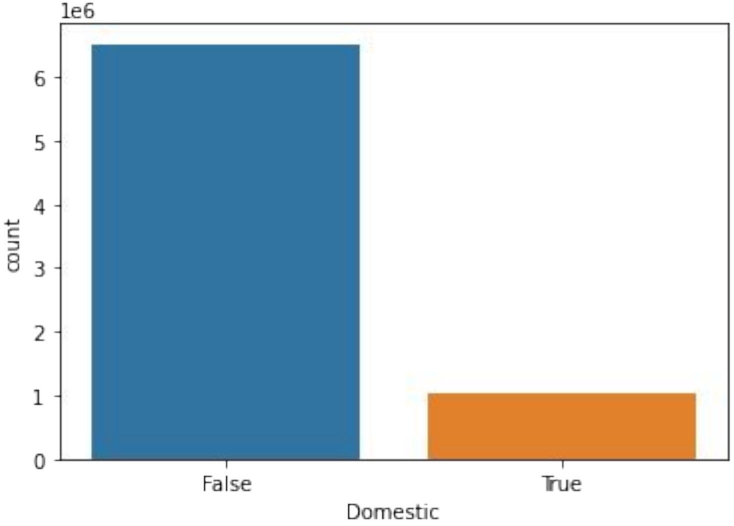
- Data Source - Chicago Data Portal
- Data Dimensions
    - ~7.5 million samples (crimes)
    - 23 features
        - 6 numerical Variables (Latitude, Longitude etc.)
        - 2 Boolean Variables (Arrest Made? Domestic Violence?)
        - 14 Categorical Variables
        - 1 Time-variable
- The columns represent major details of a crime such as type of crime, timestamp, location, whether an arrest was made, police beat information etc.
- The type of crime is our target label and has 36 unique classes.

Slide Contribution: Yash K, Yash B, Mayank A, Deb G

Exploratory Data Analysis

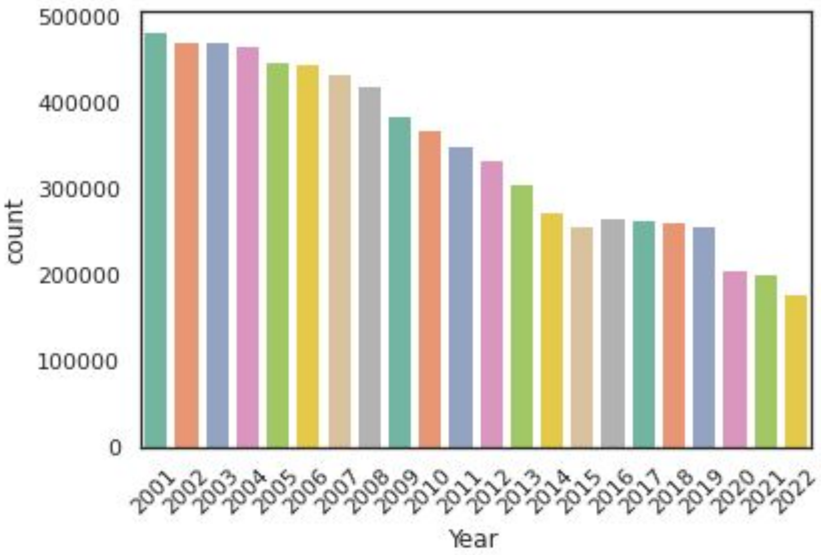
Less Arrests and Lower Domestic Crime...





Slide Contribution: Yash K, Yash B, Mayank A, Deb G

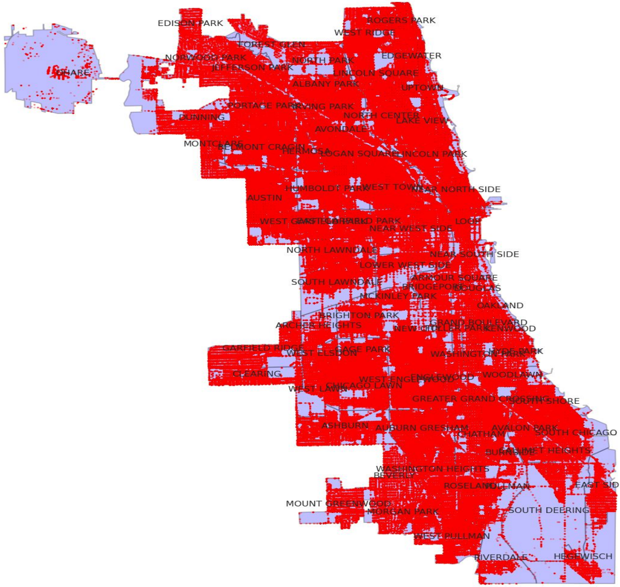
Volume of crime has a decreasing trend YoY



Slide Contribution: Yash K, Yash B, Mayank A, Deb G

Crime Hotspots in Chicago

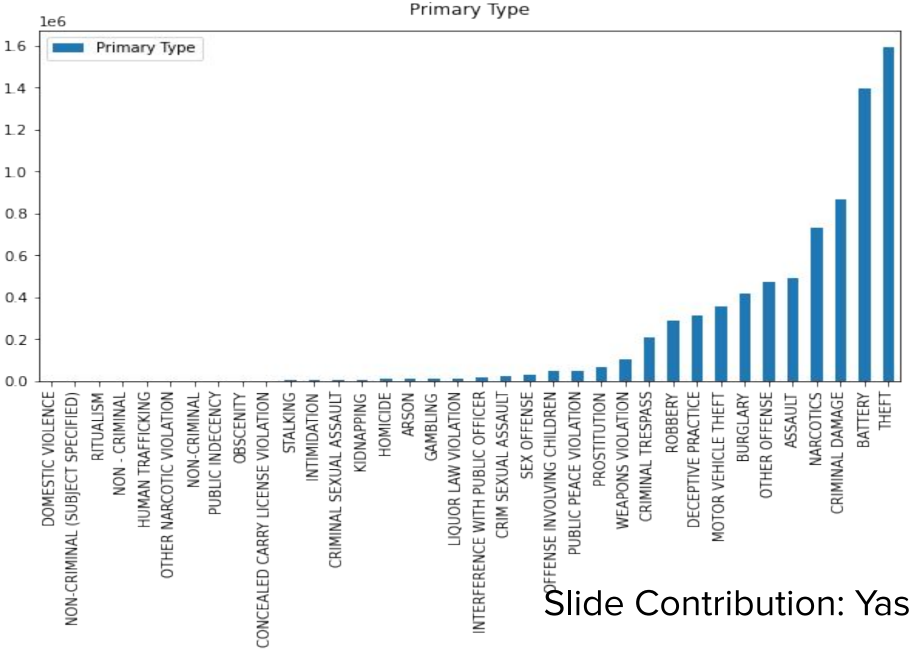
We can observe that a few neighbourhoods towards the south and west are completely devoid of crime while most other neighbourhoods have some form of criminal activity



Slide Contribution: Yash K, Yash B, Mayank A, Deb G

Class Imbalance

We can observe that the occurrence of crimes is skewed towards Theft, Battery, Narcotics and Assault leading to class imbalance.



Slide Contribution: Yash K, Yash B, Mayank A, Deb G

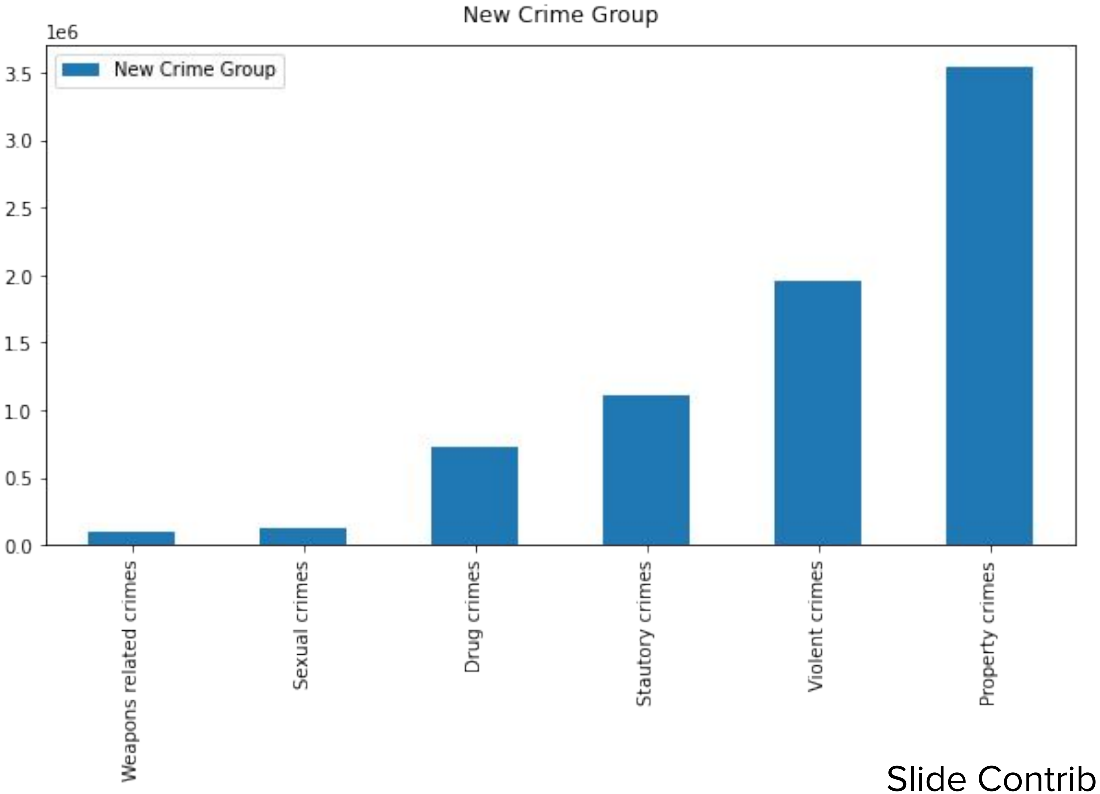
Challenges and Remedies

Challenges and Remedies

Challenges	**Remedies**
Processing the entire dataset was proving to be difficult as we didn’t possess enough compute resources to perform operations on 7.5 million samples.	Filtered the data to last 5 years to reduce the number of rows and decrease processing times without impacting RAM space
Due to acute class imbalance, logistic regression wasn’t able to predict all the classes	Grouped the 36 classes into 6 super classes based on commonalities in crime types.

Slide Contribution: Yash K, Yash B, Mayank A, Deb G

Target variable distribution based on the new classes



The new class of 6 major types of crimes is still imbalanced, but this distribution is much better than the 36 classes in the previous target variable.

Slide Contribution: Yash K, Yash B, Mayank A, Deb G

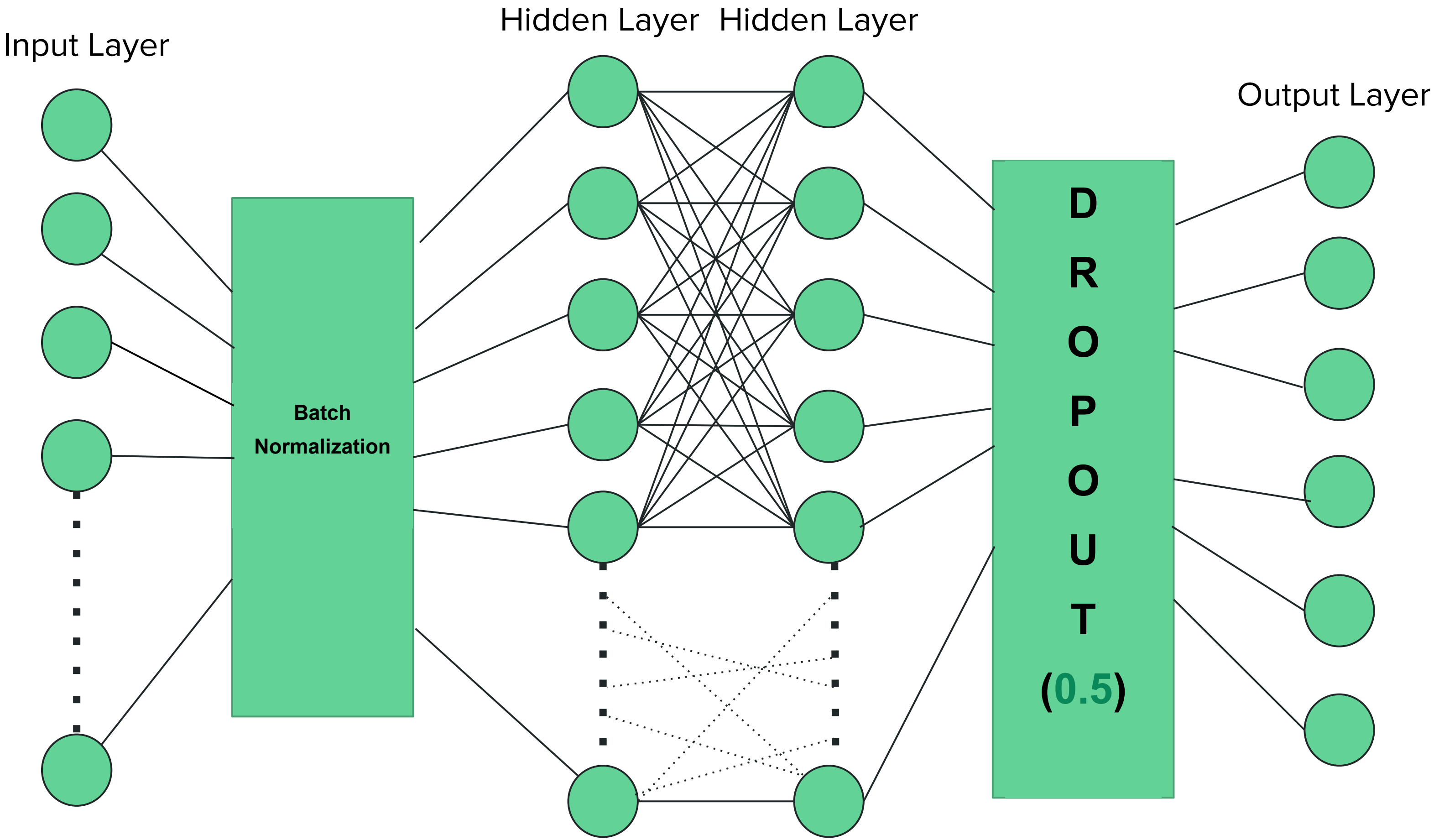
Data Pre-Processing Steps

- Converted Data to pickle file for faster load time.
- Converted DateTime column to TimeStamp for ease of manipulation.
- Extracted Year column from TimeStamp.
- Filtered Data for the last 5 years to decrease processing time and improve data handling capacity.
- Dropped rows which had Missing values in Location features.

Slide Contribution: Yash K, Yash B, Mayank A, Deb G

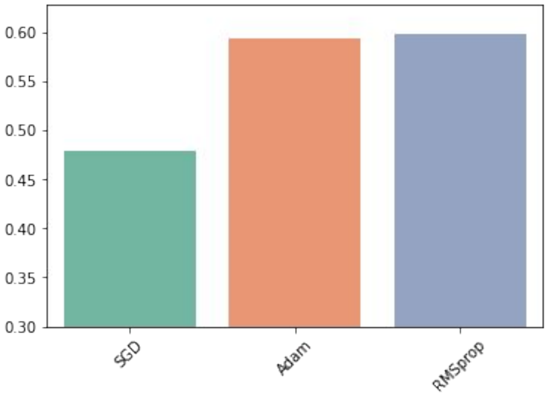
Model Architecture

Slide Contribution: Yash K, Yash B, Mayank A, Deb G

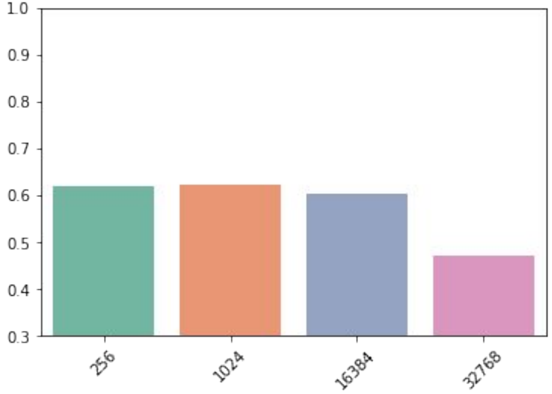


Model Selection - Hyper-parameter Tuning

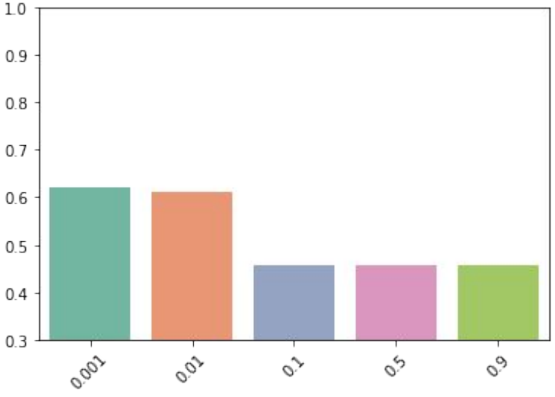
Optimizer



Batch Size



Learning Rate

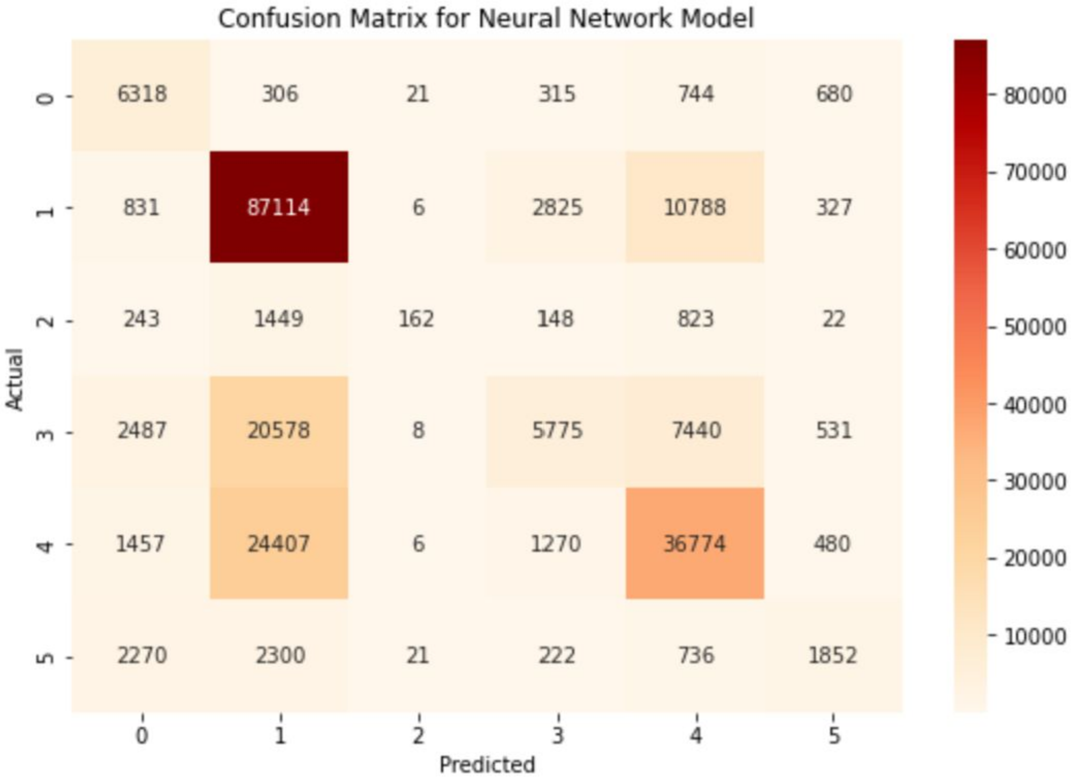


We can observe that the RMSprop Optimizer, a batch size of 1024 and a learning rate of 0.001 gives the best accuracy

Slide Contribution: Yash K, Yash B, Mayank A, Deb G

Results and Conclusions

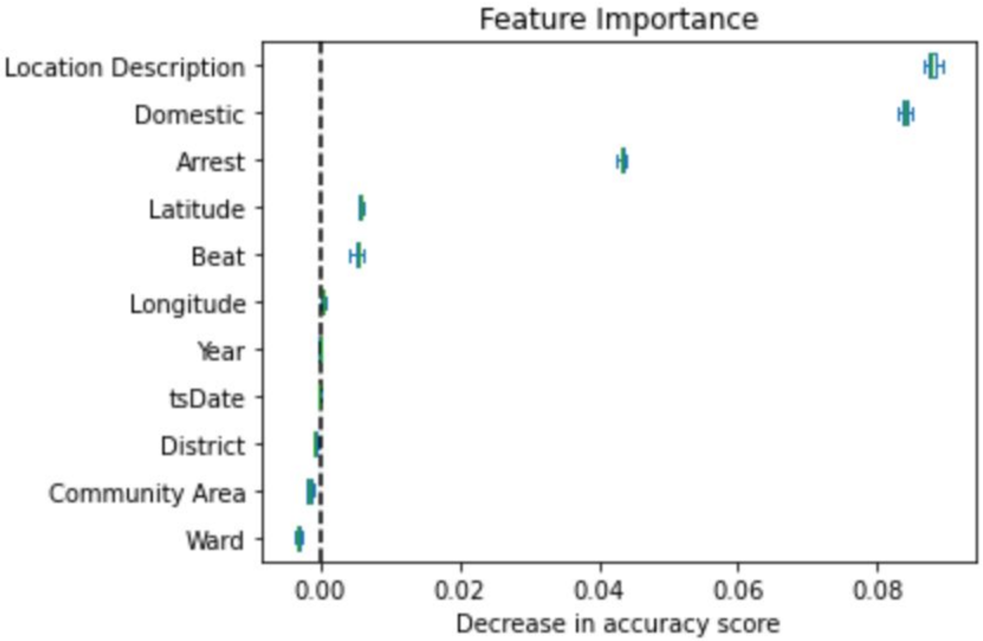
Metric	Logistic	Random Forest	Neural Network
Accuracy	0.45	0.42	0.62
Avg. Precision	0.07	0.33	0.58
Avg. Recall	0.16	0.41	0.44
F1 Score	0.10	0.31	0.43



Slide Contribution: Yash K, Yash B, Mayank A, Deb G

Results and Conclusions

- Our model has an accuracy of 62%
- Our model performs better than Logistic regression and Random forests models.
- Location Description, Domestic and Arrest are important features



Slide Contribution: Yash K, Yash B, Mayank A, Deb G

References

1. Crimes - 2001 to Present | City of Chicago | Data Portal
2. Chicago, IL, Crime Rate & Safety | U.S. News Best Places
3. GeoPandas Mapping Chicago Crimes | Kaggle
4. Predicting Crime and Other Uses of Neural Networks in Police Decision Making - PMC

THANK YOU FOR LISTENING!

Slide Contribution: Yash K, Yash B, Mayank A, Deb G