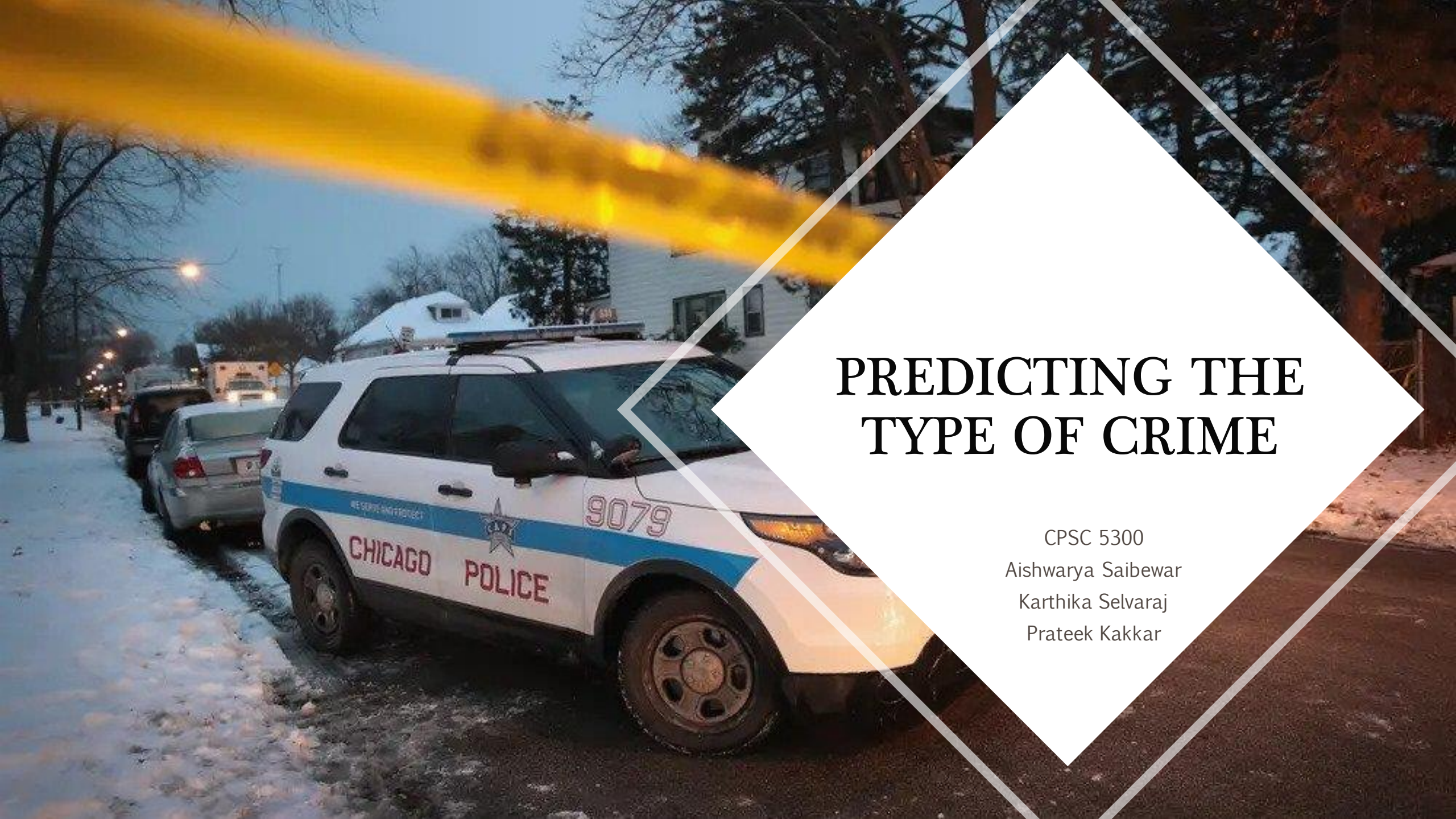


PREDICTING THE TYPE OF CRIME

CPSC 5300
Aishwarya Saibewar
Karthika Selvaraj
Prateek Kakkar

Roadmap

Problem Statement

- Predict what kind of crime is going to occur in Chicago given the selected features.

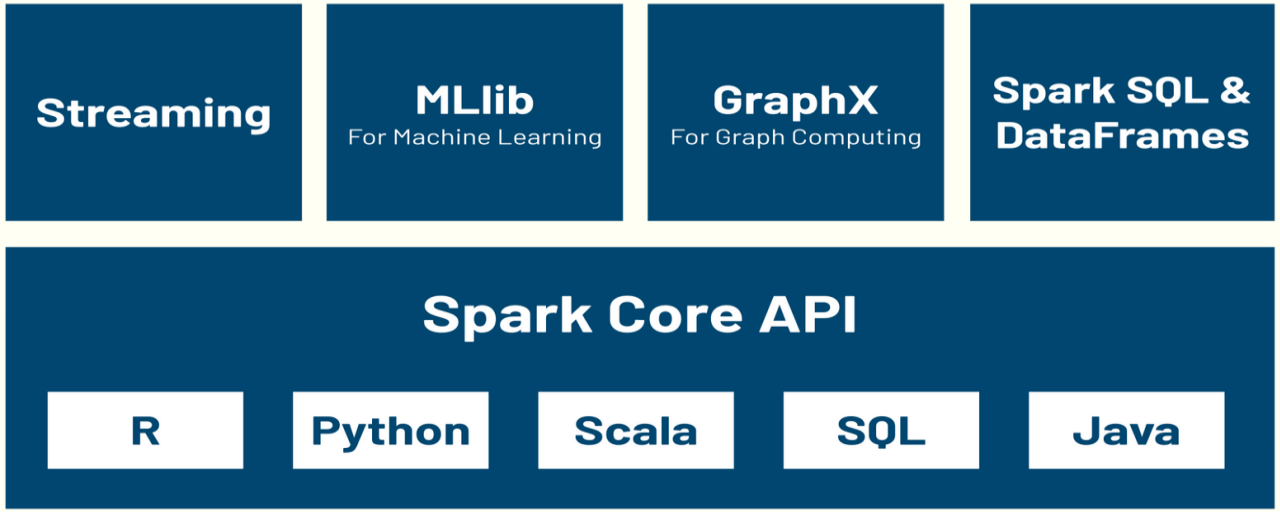
Location Description	Beat	Community Area	Day of the Week
Arrest	District	FBI Code	Year
Domestic	Ward	Hour	Month

- This dataset contains 7.55M records and 22 variables which summarize the reported crimes that occurred in the City of Chicago from 2001 to the present

- Why it is important?
  - Create a surveillance system for CPD to supposedly predict violent crime.
  - Help form violence reduction strategies that may help reduce gun violence.

Tools – Spark SQL

- Spark SQL is a Spark module for structured data processing.
- It provides a programming abstraction called DataFrames and can also act as a distributed SQL query engine.
- Provides powerful integration with the rest of the Spark ecosystem.
- Brings native support for SQL to Spark and streamlines the process of querying data stored both in RDDs and in external sources.



Tools – Spark MLlib

- Spark's library for machine learning is called MLlib.
- It contains higher-level API built on top of DataFrames for constructing ML pipelines.
- It standardizes APIs for machine learning algorithms to make it easier to combine multiple algorithms into a single pipeline, or workflow.
- In this library to create an ML model the basics concepts are:
  - DataFrame
  - Transformer
  - Estimator
  - Pipeline
  - Parameter

Tools – Spark MLlib

DataFrame : ML dataset which can hold a variety of data types .

Transformer : Transforms a DataFrame with features into a DataFrame with predictions

Estimator : Fit on a DataFrame to produce a Transformer

Pipeline : Chains multiple Transformers and Estimators together (worklfow)

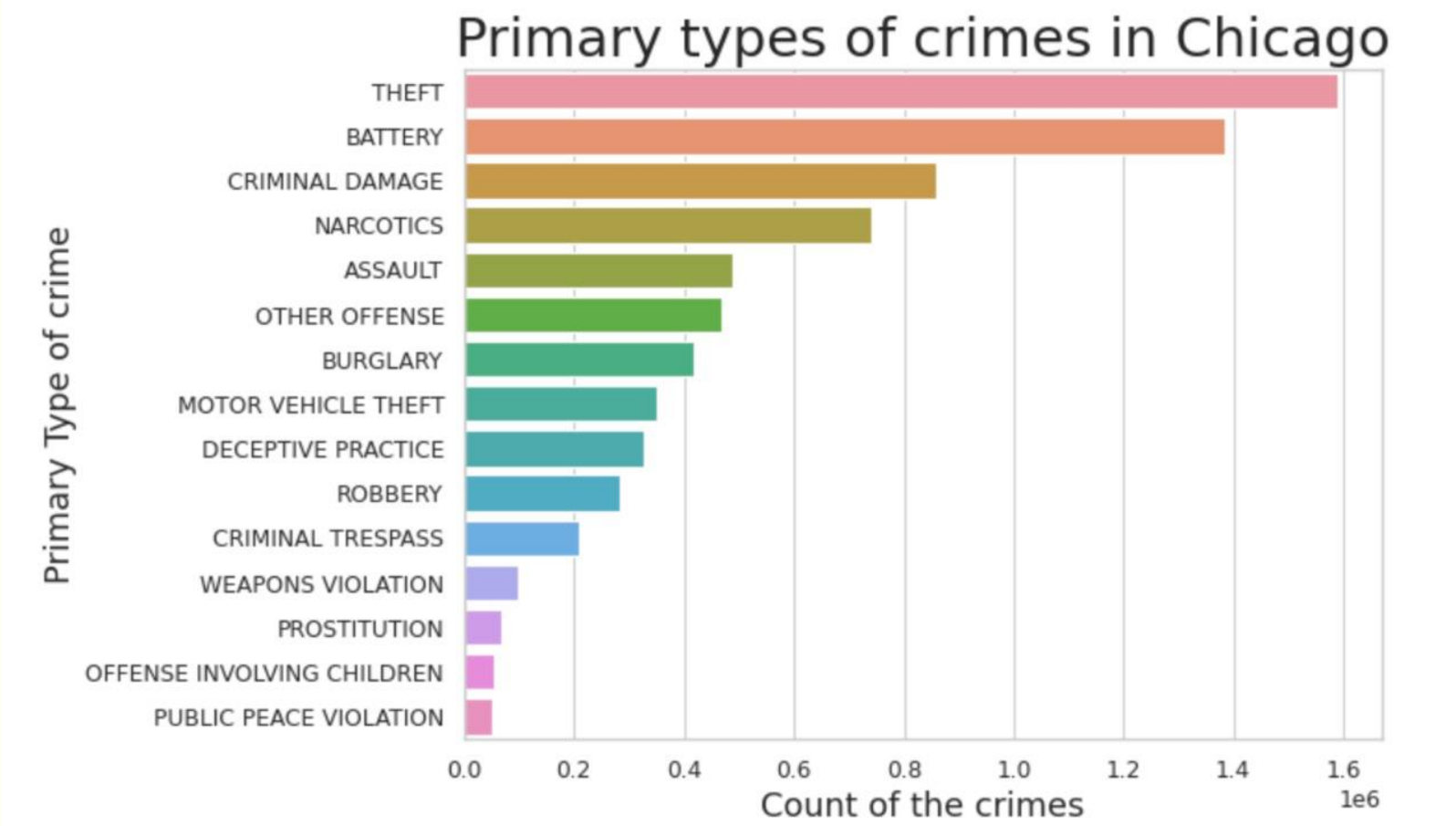
Parameter : A common API for specifying parameters of Transformers and Estimators

ML Model

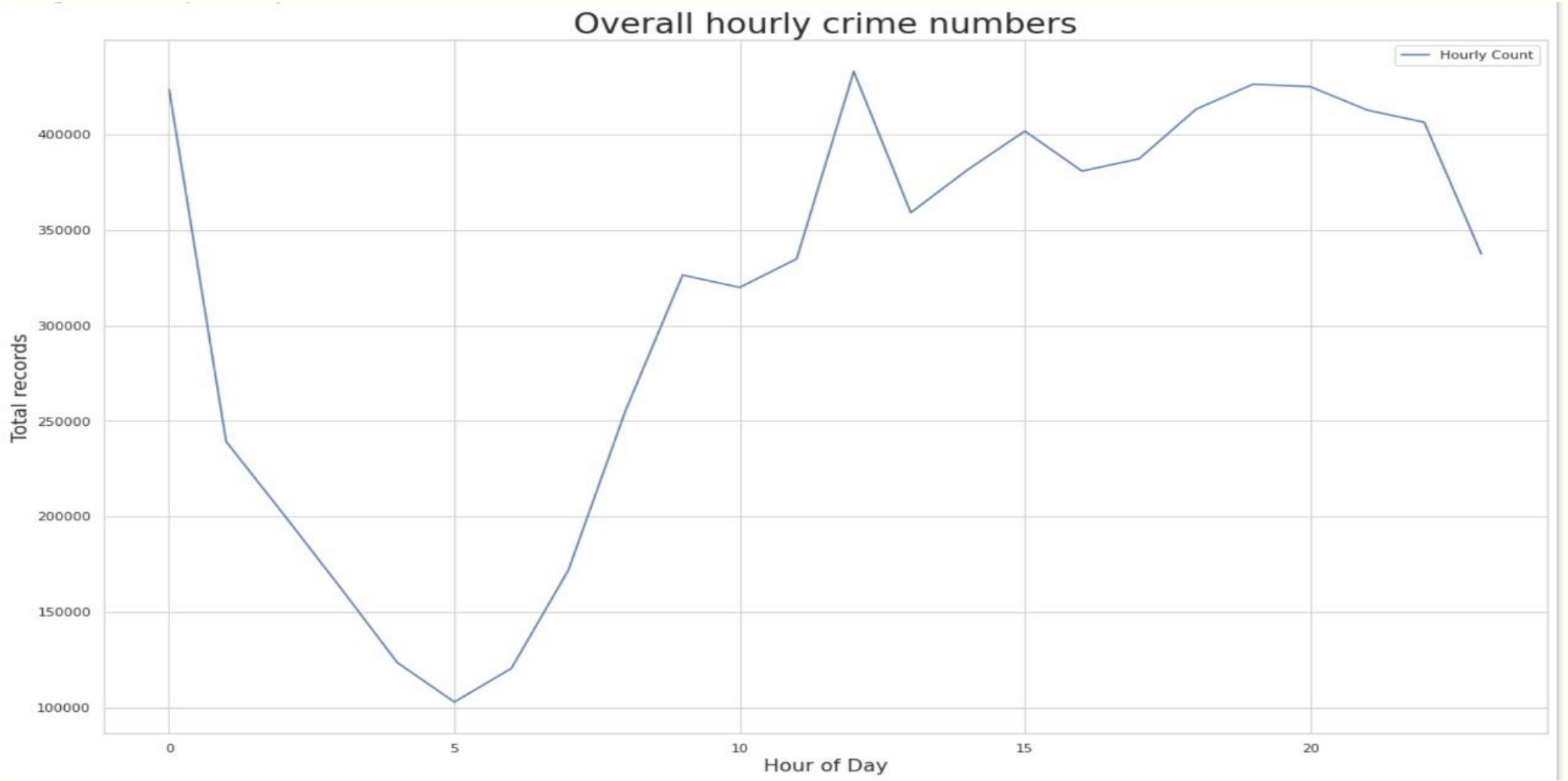
- Logistic Regression
  - Logistic regression is a popular method to predict a categorical response.
  - Predicts the probability of the outcomes.
  - Fast at classifying unknown records.
  - Provides inference about the importance of each feature.

CRIME DATA VISUALIZATION

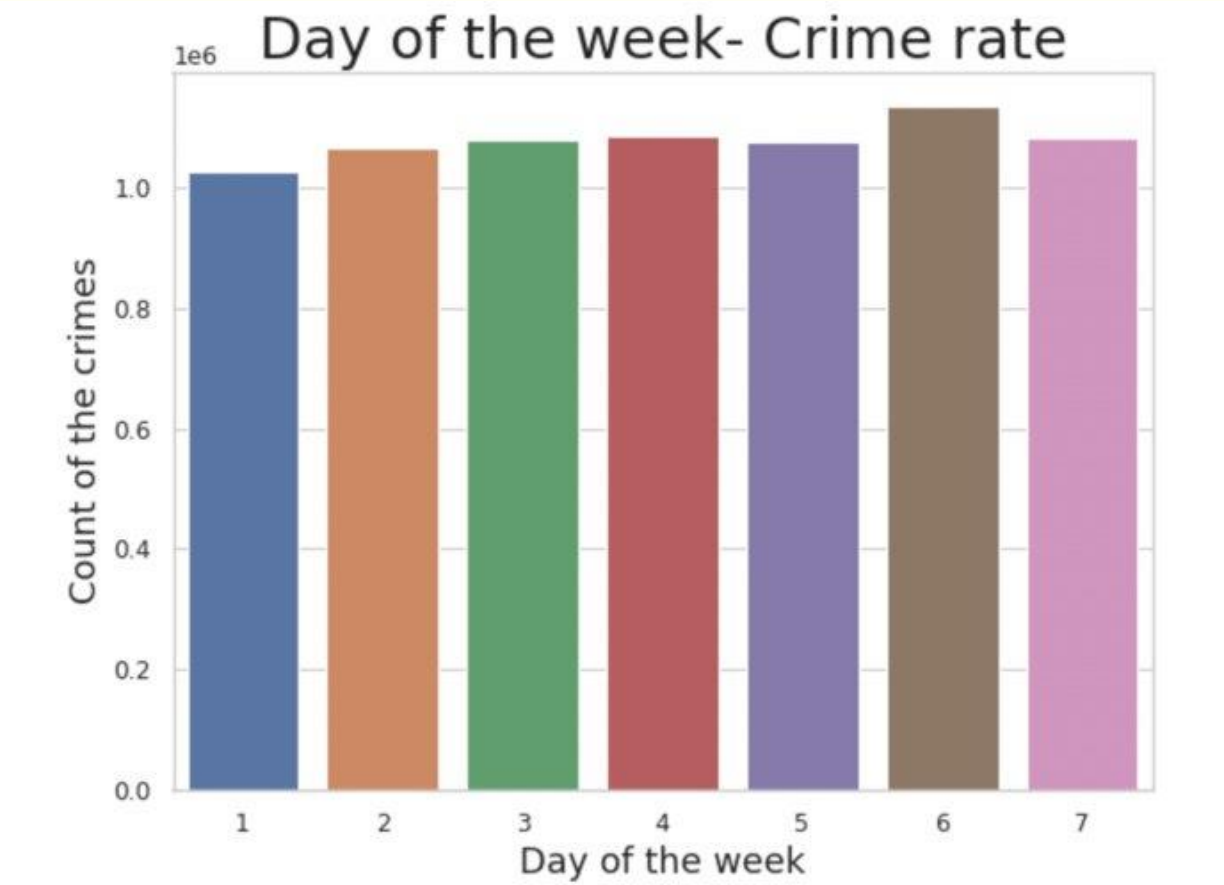
MOST COMMITTED CRIMES IN CHICAGO



TRENDS IN CRIME RATE BY TIME OF A DAY

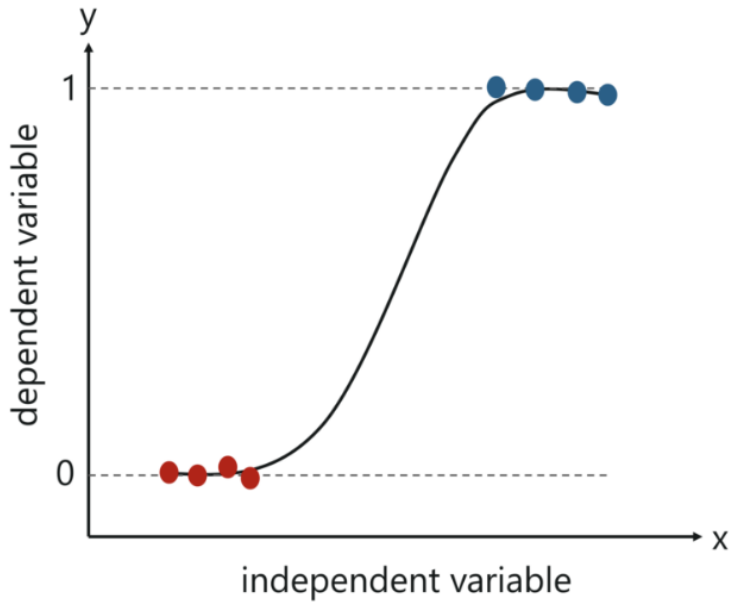


TRENDS OF CRIME OVER A WEEK



MODEL CREATION

Logistic Regression



- Logistic regression is a method used to predict a dependent variable , given a set of independent variables, such that the dependent variable is categorical.
- Y is the probability of an event to happen which, we are trying to predict.
- X1, X2.. Are the independent variables which determine the occurrence of an event

$$\log\left(\frac{Y}{1-Y}\right) = C + B1X1 + B2X2 + \ldots.$$

Binomial vs Multinomial logistic regression

- **Binomial Logistic Regression**: Standard logistic regression that predicts a binomial probability (i.e., for two classes) for each input example.
- **Multinomial Logistic Regression**: Modified version of logistic regression that predicts a multinomial probability (i.e., more than two classes) for each input example.

Steps in creation of a Logistic Regression Model

- Map the string column of labels to ML columns of label indices using String Indexer.
- Merge multiple columns into a vector column using Vector assembler.
- Split the data into train and test.
- Build the logistic regression model by specifying the response variable, predictor variables and the parameter family as multinomial.
- Fit the model on a training dataset.
- Check the accuracy of the model on training and test datasets.

Results:

The model predicted with an accuracy of 62%

DEMO

CONCLUSION

- The most committed crime in Chicago was visualized and the primary type of crime is found to be theft followed by criminal damage and narcotics
- The lowest crime rate is observed during the early hours, and it gradually increases to the maximum in the afternoon
- The crimes are distributed equally for the days of the week, but a little low on Monday and little high on Saturday.
- The Logistic regression model predicted the primary type of crime with an accuracy of nearly 62%
- Future analysis would be on using the Tree based model in predicting the primary type of crime.

THANK YOU!