Final Project Proposal

Advance Data Science & Architecture

Team 9 - Aashri Tandon, Pragati Shaw, Sarthak Agarwal

Dataset: Chicago Crime data

Dataset Description:

Source: City of Chicago link

Data Size: 1.4 gigabytes

Columns: Dataset has 22 columns.

Columns
ID
Case Number
Date
Block
IUCR
Primary Type
Description
Location Description
Arrest
Domestic
Beat
District
Ward
Community Area
FBI Code
X Coordinate
Y Coordinate
Year
Updated On
Latitude
Longitude
Location

Data Download and Preprocessing:

We will download the data from the city of Chicago official website.
https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data

- Perform Missing data analysis. Check the percentage of missing values and based on frequency distribution, handle them.
- Perform Feature Engineering. Check for correlation and do engineer features as needed.

Exploratory Data Analysis/ Data Visualization

Exploratory data analysis will serve two purpose. Firstly, we will learn insights about the data and secondly we will display the best analysis that will be beneficial to our end user in the web application.

- Perform univariate and bivariate data analysis to get insights about the data.
- Plot data visualization. E.g.
    - How has crime changed over the years?
    - Which areas of the city have evolved over this time span?

ArcGIS:

We will implement ArcGIS using its python API for doing geo spatial analytics and showing interesting metrics to the user. Like comparison of neighborhoods crime rate and its effects of one community on other, etc.

Machine Learning Component:

Clustering: Divide the regions in Chicago into different clusters based on the crime rate.

Prediction: Then, build a prediction model per cluster that will predict the probability of the next crime.

Choose the best prediction model from Linear Regression, Random forest and SVM.

End User Web Application

The user will enter a location in a dialog box on the web page. That area will get highlighted on the map and our application will display various insights about that place that will be a result of our data exploration. It will also contain some dynamic graphs that will display location based insights like

Next, it will display the probability score of the crime happening in the future from our prediction model.

Infrastructure

Python – Data processing and Machine Learning.

Docker – For easy distribution and submission.

Java – Web application.

Microsoft Azure – Machine learning Rest API

Tableau/Plotly – Data Visualization

ArcGIS – Data Exploration and Analysis