Proposal on the crime data analysis

Commented [Office1]: The choice of color and the contrast in font size (compared with the size you use lower on this page)

Date: November 1, 2019

From:

Madhuri Yadav, Graduate Student at GWU

Tanaya Kavathekar, Graduate Student at GWU

Table of contents

Purpose

The purpose of this proposal is to present an overview and a plan for an academic project.

Summary

In the era of big data and emerging machine learning algorithms, there is a growing interest in studying crime records data to take proactive measures instead of reactive. The goal of a crime department is not just to reduce the number of crimes but to avoid crimes. And hence, there is a need for mining data and applying machine learning on crime data. The objective of this project is to identify and analyze various factors that affect crime in India. Although several factors play a significant role in determining the rate and trend of crime, due to the limitations of data and time, we will be focusing on only a few factors such as Location, Time, Age, Sex, Income Level, etc.

Introduction

Crime can be defined as an action or omission that results in an offense that is punishable by the law. Any criminal activity causes loss or damage to life, wealth, mental health. Data is the new weapon to fight against crime. Analysis of the crime-data is very crucial for crime investigations.

The overall crime rates have been decreasing over the years throughout the globe. However, according to Ansari, Verma & Dadkhah (2015), the rate varies depending on the type and the region of crime. For example, crime in burglary and theft have reduced over the years but crimes related to drugs have increased. Similarly, the crime rate varies considerably across the regions for example, murder rates have decreased in Europe but have increased in the USA.

A lot of research has been conducted to understand the trends of criminal records across European and North American countries (James 2018; Aebi and Jehle, 2018; Lewis, Barclay, De Cavarlay, Costa and Smit, 2004; Bjs.gov, 2019). However, there are not many common measures to scale the analysis of Indian crime trends.

Problem Statement

The objective of the project is to identify and analyze several factors that may affect several crimes in India. This will be useful to devise optimal strategies to prevent crimes.

Dataset

We are going to use the data set collected by open government data platform India. The Dataset contains 9 years of data from 2001. It is a state-level dataset. The data set requires preprocessing to remove invalid values and requires data integration since data is distributed among multiple columns. The shortcoming in this dataset is there are a lot of missing data when it comes to the crimes reported by the local police department. And, since we have missing crime reports for the

crimes which were not reported in the police department. Here is the link to the dataset being used for our purpose.

Proposed Tasks

Task 1: Defining the problem

We have already defined the problem and scope of the problem after preliminary research on different categories of crimes in India, variation in crime across time-period. The current scope of the problem will be limited to the available data at the source.

Task 2: Data pre-processing

The data from the identified source cannot be used as-is as there could be a possibility of inaccurate data. Before building a machine learning model, it is important to perform certain data preprocessing steps to avoid misleading results. Those preprocessing steps are:

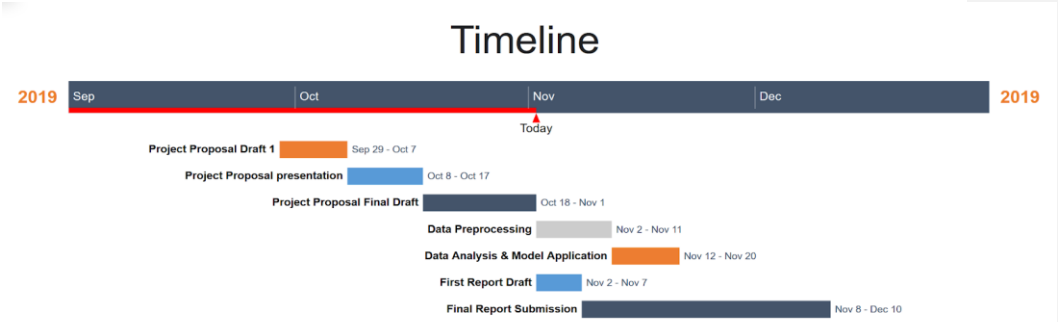
1. Understanding the data columns
2. Missing value identification and treatment
3. Outlier detection and treatment
4. Data statistics
5. Typically, at the end of these steps data is cleaned and ready to use for analysis

Task 3: Identifying the relationship between features and a dependent variable

In this case, the dependent variable is the number of crimes. There are various statistical techniques namely Pearson correlation, ANalysis Of VAriance (ANOVA), chi-square test which can be used to find a relationship between variables. For these algorithms, we have to define the null hypothesis and the alternative hypothesis. The null hypothesis is often what we try to prove or disprove. In our case, the null hypothesis for a test will be there is a relationship between variables.

At the end of the above-mentioned steps, we would have successfully identified a set of features that affect several crimes.

Schedule



References

Aebi, M. and Jehle, J. (2018). Introduction to the Special Issue on Crime and Criminal Justice in Europe. *European Journal on Criminal Policy and Research*, 24(1), p 3-6.

Ansari, S., Verma, A., & Dadkhah, K. (2015). Crime Rates in India. *International Criminal Justice Review*, *25*(4), 318-336. doi: 10.1177/1057567715596047

Bjs.gov. (2019). *Bureau of Justice Statistics (BJS) - All Data Analysis Tools*. [online] Available at: https://www.bjs.gov/index.cfm?ty=daa.

James, N. (2019). Recent Violent Crime Trends in the United States. Retrieved from https://fas.org/sgp/crs/misc/R45236.pdf

Lewis, C., Barclay, G., De Cavarlay, B., Costa, M., & Smit, P. (2004). Crime Trends in the EU. *European Journal On Criminal Policy And Research*, *10*(2-3), 187-223. doi: 10.1007/s10610-004-2569-y

Nationmaster.com. (2019). *European Union vs United States: Crime Facts and Stats*. [online] Available at: http://www.nationmaster.com/country-info/compare/European-Union/United-States/Crime.

Ucrdatatool.gov. (2019). *Uniform Crime Reporting Statistics*. [online] Available at: https://www.ucrdatatool.gov/