Capstone Project
Safe or Unsafe?

Index

Introduction

Growing up in India, my parents have always had a hard time letting me leave the house at night and I had a strict curfew at 8pm. Yet somehow at the age of 18, my parent allowed me to come to USA to pursue higher education, and they live in constant fear. I feel if there was a way of knowing whether the neighborhood I'm about to enter is safe at that time, it would greatly assuage their fears.
So for this project, I would like to create a model to predict whether a neighborhood is safe at a particular time of the day. The location chosen for this project is Oakland, since I frequently travel to Oakland and use the BART. I would like to use a prediction model, but haven't yet decided on a particular algorithm.
This app is targeted towards anybody who needs to check whether the area they are about to enter is safe during that time slot, especially students travelling alone. Ideally, I would have liked to create a general model which can be applied to any city but the form of data collection differs and each dataset would require a unique approach to data cleaning.

Data

I will be using the Geopy package to obtain the location of the neighborhoods in Oakland. The information extracted will also be used to create maps (using Folium package). The data about the crimes committed will be taken from the https://data.oaklandnet.com/Public-Safety/CrimeWatch-Maps-Past-90-Days/ym6k-rx7a site. I will be using the csv file dataset and making additional calls to the SODA API if needed. This contains the information on the crimes that have occured in the past 90 days so the information is not outdated. The data contains the date and time of the crime, description, and the address of the location of the crime. I will be able to obtain the latitude and longitude using that address. The columns of the csv file are as follows:

Column Name	**Description**	**Example**
CRIMETYPE	Plain text (a brief description of the crime)	'Weapon', 'Domestic Violence'
DATETIME	The date and time of the crime (yyyy-mm-dd-hh-MM-ss-p)	04/02/2020 07:42:00 AM
CASENUMBER	An ID given to the case (string)	20-017349
DESCRIPTION	Contains more details about the crime	CARRY CONCEALED WEAPON IN VEHICLE
POLICEBEAT	Indicates the 'beat' where the crime occurred (beat is the smallest geographic area)	19X
ADDRESS	The street address of the crime	1500 E 12TH ST
CITY	Oakland for all rows	
STATE	CA for all rows(since the city is the same)	
LOCATION	The address along with the latitude and longitude of the location in string format	1500 E 12TH ST\nOakland, CA\n(37.787943, -122.94586)

Methodology

Data Cleaning

For cleaning the data, I had a particularly difficult time since this was my first time doing a project without any scaffolding. There were no missing values, but the format of the table made it difficult to use. For instance, the address was a string containing the address followed by the coordinates so I had to separate that to extract the coordinates. Even after obtaining the latitude and longitude, I was unable to find a package that would reverse-geocode the locations for almost 16,000 rows so I decided to split the city into Areas. Each of these areas had a list of police beats within them so this made it easier to split the city.

Another thing I noticed, was that amongst the type of crimes committed, there were certain categories which wouldn't affect me or anyone who was just about to step into the neighborhood for a small period (Vandalism). So I decided to segregate the columns into 0 and 1 according to the type of crime committed and a SAFETY column was added to the dataframe which would eventually become our target column.

The dataframe provided to us did have a time column but I was unable to find anything which would help divide the rows using the time. I intended on checking if the hour of the day affects the safety of an area so it was a crucial step. I decided to convert the time into 1-dimensional i.e. I converted the time into minutes. For example: 7:30pm -> 19:30 -> 19*60 + 30 = 1170. Since I wanted 6 slots of 4 hours each, I simply used the cut function to split 0-1440 into slots of 240. This made my job much easier and I was able to plot this with ease.

Methodology

Model

I used the process of elimination to decide upon a model. Since my target column was not continuous variables, I was able to eliminate Regression. My features were not continuous either so I could not use Logistic Regression. Ultimately, I decided to go with a Decision Tree.

In order to train and test my model I used the same dataset, with 30% of the dataset used for testing. I used the train_test_split() function to split my dataset.

Of all the columns, I decided to use the month, time and area number as features for my decision tree. The target variable was the safety.

Ultimately, I could only get to 62% accuracy, after which I tried tweaking the features, and keeping 1 feature variable and the others constant. But after a lot of trial and error this was the highest accuracy I could achieve with the given information.

Results and Discussion

During the data analysis we found that there were certain areas which had a greater crime rate than the others and this seems to match the information available on the internet about the neighborhoods in Oakland (Rockridge is the safest). We also found that while the days of the week did not seem to affect the numbers as much, the time slots really showed a stark difference in the crime rates. Surprisingly, we find that 4pm-8pm is the most common time for a crime to take place when that is actually the time when the public, students and workers complete their work or classes for the day and are heading home or heading out with friends. Unfortunately this was a big dataset due to which I had a hard time using reverse-geocode to find the postal codes of the scenes of the crime but instead I chose to Police Beats to divide the city into areas. These, incidentally are also the BART stations around Oakland. So this model would be even more helpful for those getting off at one of these stations.

Conclusion

In hindsight, I should have used a smaller database. I intend to continue improving this model, maybe reduce the number of rows and get more data from the internet which affects the crime rate. One of the reasons for the low accuracy could be that there are many other factors which have not been taken into taken into consideration. Lastly, even if I am able to build a model with high accuracy, it would still be impossible to predict if an incident happening across the country could affect the crime rate here.