Business Understanding Document

Vivek I(IM2013017)
Aravind K(IMT2013019)
Nikhilesh P(IMT2013031)
Arun Prasad(IMT2013036)

Introduction

As time proceeds the number of crimes in California are increasing. So, it's better to know which city to live in if we need to be safe. Any type of crime which leads to death of people is more dangerous. So, if we know the mortality rate of each type of crime, we can know which crime to be taken more seriously and to be far against. In fact, these kind of analysis is useful not only for California but for any state. These are the basic motivations of this analysis project we have chosen.

Since, we are taking the data of California state. This analysis would be helpful for any individual who wants to live in California state. Since, we are also providing mortality rate of each crime, it would be useful to the police department to concentrate on that particular crime which has more mortality rate.

So, our problem mostly falls under the category of public safety by crime analysis.

Problem Description

Our analysis project aims on finding the safest city (or) town in California state. Precisely speaking, our aim is to give rank to each city (or) town based on the safety in that city (or) town. One more aim of the project is to find the mortality rate of each type of crime.

Pre-requisites of the project are the statistics about number of crimes in each city (or) town along with the type of crime. (This is the data we are using for doing the analysis)

So, the two goals are:

-Giving a rank (from 1 to number of cities) to each city based on safetiness in that city.

-Giving a mortality rate for each type of crime in California state on a scale of 1 to 10 where 1 indicates least mortality rate and 10 indicates highest mortality rate.

The term 'safety' may have different notion for different people. So, while doing analysis, we give a definions for safety and give the rank for each city based on the give definition.

Resource Inventory

The data we are using is the CRIME in the united states 2013 data from ucr.fbi.gov website.

The data available is in the tabular form which contains the number of cities and towns of California along with the number of crimes of each type in that city.

Data dictionary for the data:

- City- Name of the city (or) town
- Population- Population of that city
- Violent crime, Murder and no negligent manslaughter, Rape, Robbery, Aggravated assault, Property crime, Burglary, Larceny-theft, Motor vehicle theft, Arson columns shows the number of crimes of that type.

Translating the problem into Data mining goals:

For safest city:

-Give rank to each type of crime based on its safety (as per the defintion we considered)

-Based on population in each city, number of crimes of each type in a city, weightage given to each crime based on safety involved in that crime, we give rank to each city based on safety in that city

For mortality rate:

-Give a rank to each crime based on its mortality rate

-Based on the population of each city, number of crimes of each type in a city, mortality rate of each type of crime mortality rate of each type of crime on an aggregate level in the California state is given.

Data mining Techniques:

The main objective of the project is to determine the mortality rate caused by each type of crime and to determine the safest cities or towns in California to live in. In order to achieve these requirements, we use few data mining strategies as described below;

(I) To get the mortality rate caused by each type of crime we incur descriptive analytics, that is by taking the number of deaths caused by that crime, total number of deaths caused in that city and the total population of that city into consideration, we allocate a factor to that crime from the scale of one to ten which is hence determined by summarizing the available data.

(II) To generate the list of cities in the order of their safety of living, we first determine the ratio of the number of victims of a particular crime to the total population in that city and allocate a weightage for each of the crimes in the dataset based on the previously determined mortality rates and our definition of safety. Based on the factors that we are going to finalize in the further run and their corresponding ratios and weightages we classify these cities or towns into whether they are safe to live or not based on a threshold and determine their order using the resultant values.

Data Mining Success Criterion:

- Calculate specificity, sensitivity and accuracy for the model to analyse the results
- We are giving a score to each and every city in the California state on a scale of 1-10. Any city with a score more than 6.5 would be regarded as a safe city.
- Successful deployment of model results is part of data mining process.

Project Plan

We are planning to follow the below time table for our project (total of 9 weeks' duration)

Phase	Time
Business Understanding	1 week
Data Understanding	2 weeks
Data Preparation	3 weeks
Modelling	2 weeks
Evaluation	1 week