Big Data Analytics - Team Project Proposal

Part 1. General Project Information

Names of team members: Carlos Petricioli (cpa253), Valerie Angulo (vaa238), Varsha Muralidharan (vm1370)

Project Title: Predicting crime using taxi trips data

Project Description: ***(Write one paragraph to describe what this analytic will do.)***

Understanding and predicting crime is a crucial task in any mayor city. The objective is to understand crime rates at a granular level with the idea that people behave according on how secure they feel and this fact impact the way they travel. It might be that people prefer to take a taxi versus other options depending on their own perception of crime in their current location. This work will analyze taxi and crime data on a case level.

This is a modern approach that will complement the use of demographics and geographical variables commonly used to predict crime. Global Positioning System (GPS) data on taxi rides provide useful information that can be directly related to crime at a block level. There is enough data to make this analysis possible.

This work will be limited to the City of New York.

Who is a typical user of your application:

The scientific community, the citizens and demographics themselves.

What insight will you derive from the data?

The objective is to get a better sense of the way that crime relates to the use of taxis in NYC.

Describe how you will prove the goodness of your analytic, in other words, how will you verify that it is correct:

By using traditional machine learning standards such as Train/Test data, cross validation and model validations where applicable.

Big Data Analytics - Team Project Proposal

Part 2. Data Source Information

Name of Data Source 1: Taxi data from the TLC.

Data Source Description: The yellow and green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. It covers years from 2009 to June 2017.

Data Collection Frequency	**Data Size**	**Data Frequency** • Every time it gets updated, which is about once a month.
Batch (multiple non near-realtimecollections)	250 GB	• Will you collect a batch of data periodically or just once (static)? Every month. • How much data that will be collected at each interval? Once a month data grows about 1GB.

Big Data Analytics - Team Project Proposal

Part 2. Data Source Information

Name of Data Source 2:

• NYPD Complaint Data Historic • NYPD Complaint Data Current YTD

Data Source Description:

This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of last year (2016).

This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) for all complete quarters so far this year (2017).

Data Collection Frequency	**Data Size**	**Data Frequency**
Batch (multiple non near-realtimecollections)	1.5 GB	• Will you collect a batch of data periodically or just once (static)? Once every quarter • How much data that will be collected at each interval? If the data will be collected periodically, how often will you collect it and what is the volume of data that will be collected at each interval? Once every quarter grows about 100MB

Big Data Analytics - Team Project Proposal

Part 2. Data Source Information

Name of Data Source 3: NYC Weather stations data from NOAA (Central Park, JFK and Laguardia)

Data Source Description:

The Integrated Surface Database (ISD) consists of global hourly ansynoptic observations compiled from numerous sources into a single common ASCII format and common data model. ISD's complete history of hour-by-hour readings for one user-specified weather station

Data Collection Frequency	Data Size	Data Frequency
Static (one timecollection)	200 MB	**If *not* realtime data:** • Will you collect a batch of data periodically or just once (static)? □ Just once