Spring 2024 Big Data: CSGY 6513-D

Name: Jiaxin Dong, Chenchen Guo, Mohammed Zakriah Ibrahim
NetID: jd5226, cg4421, mi2471

Big Data Project Proposal

Abstract

This project proposal aims to conduct an extensive big data analysis of traffic crashes within the City of Chicago, leveraging the comprehensive crash data made available through the City's data portal. This dataset encompasses detailed reports of each traffic crash on city streets within Chicago limits and under the jurisdiction of the Chicago Police Department (CPD). It originates from the electronic crash reporting system (E-Crash) at CPD and meticulously excludes any personally identifiable information to ensure privacy. The records, updated when a crash report is finalized or amended, cover data from certain police districts starting in 2015, with citywide data available from September 2017 onwards.

Statement

Our analysis will focus on a multitude of factors, including but not limited to, street and weather conditions, posted speed limits, and the nature of each crash, to identify patterns and insights that can enhance road safety and policy formulation. Notably, this dataset differentiates between self-reported minor crashes at police districts and those recorded at the scene by responding officers, providing a unique dataset to analyze. Our project will also take into consideration the amendments made to crash reports, which can offer a dynamic view of each incident's evolving understanding over time.

Objective

Feature Engineering:
Find correlations focusing on various parameters such as street and weather conditions, posted speed limits, and crash types. This analysis aims to understand the complex dynamics contributing to traffic crashes.

Model training: Train logistic regression models or any other classification models that can help to classify traffic accident on given feature

Pattern Identification:
To identify patterns and trends in traffic crash occurrences, including temporal patterns (time of day, day of the week, seasonal variations), spatial patterns (geographical hotspots of crashes), and situational patterns (conditions under which crashes are more likely to occur).

Risk Factor Analysis: To use machine learning algorithms and statistical methods to identify and quantify the risk factors associated with traffic crashes in Chicago. This includes analyzing the impact of environmental, vehicular, and human factors on crash probabilities.

Data Privacy and Ethics Compliance: To ensure that all analyses respect the privacy of individuals and comply with relevant data protection laws and ethical guidelines. This includes the responsible handling of data, especially in light of the exclusion of personally identifiable information such as the RD_NO.

Enhancement of Road Safety Measures: To provide actionable insights that can be used to enhance road safety measures within the city. This includes suggestions for improvements in road design, traffic signalization, speed limit enforcement, and public education on road safety.

Methodology & Technology

Python 3.10+
Pandas
PySpark
Jupyter Notebook
Sklearn
Torch

Matplotlib
Seaborn

Data source:

Traffic Crashes - Crashes Chicago data
Dataset size: 418MB
Number of Record: 810k
Link:
https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if/about_data

Backup Options Please Ignore

Human vs. LLM Text Corpus 2GB

https://www.kaggle.com/datasets/starblasters8/human-vs-llm-text-corpus/data

Text file generated from human GPT3.5 and other LLM's. Can use to detect AI text generating

Bitcoin +233 Crypto Coins Prices 2GB

https://www.kaggle.com/datasets/olegshpagin/crypto-coins-prices-ohlcv
A summarization of bitcoin price by different time scale. Can use to construct regression model

1.3M Linkedin Jobs & Skills (2024) 2GB

https://www.kaggle.com/datasets/asaniczka/1-3m-linkedin-jobs-and-skills-2024/data
A list of linkdin jobs skill list in raw. Can use to practice data clean analysis and visualization

Amazon Reviews for Sentiment Analysis 514MB

https://www.kaggle.com/datasets/bittlingmayer/amazonreviews
Amazon reviews with specific prompted labeled. Can use to train RNN prediction

Traffic Crashes - Crashes Chicago data

https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if/about_data

Major Crime Indicators Open Data Toronto

https://data.torontopolice.on.ca/datasets/TorontoPS::major-crime-indicators-open-data/about