Final Project: Proposal

Project Title: Traffic Accidents Analysis and Prediction

Group Members: Zizheng Que, Xubin Wang, Lingyun Li, Shan Chen, Yangyang Jiang

1. Research questions

List 3 questions that you intend to answer (1 point)

Detail each research question you intend to answer.

1. Analyze the correlation between different factors (e.g., weather, location, time) and car accidents to identify the most significant contributors to car accidents.
2. Determine the most hazardous locations, times, and weather conditions—that is, identify specific periods, locations, and weather conditions with the highest likelihood of car accidents.
3. Employ Machine Learning algorithms on existing datasets to predict the probability of car accidents given information such as latitude, longitude, weather conditions, etc.

2. Dataset utilization

List all the datasets you intend to use (1 point)

Name and source of each dataset, brief description.

Main Dataset:

a) US Accidents (7.7 million records): A Countrywide Traffic Accident Dataset (2016 - 2023)
Link
Description: This dataset encompasses car accident records across 49 states in the USA, gathered from February 2016 to March 2023. The collection of data was facilitated through several APIs that stream traffic-related incidents. These APIs source their information from a variety of contributors, such as departments of transportation at both the state and national levels, law enforcement bodies, traffic surveillance cameras, and sensors embedded within the road infrastructure. Currently, the dataset holds about 7.7 million records of traffic accidents.

Supplementary Datasets:

b) Comprehensive Dataset of 33 Million U.S. Traffic Congestion Events (2016 - 2022) Link
Description: This is a countrywide traffic congestion dataset that covers 49 states of the USA. The congestion events data were collected from February 2016 to September 2022, using

multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by various entities, including the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks. The dataset contains approximately 33 million congestion records. We also provide a sampled version of data that includes 2 million events for easier processing and handling for those who prefer to work with a smaller amount of data.

c) A countrywide dataset of 6.2 million road construction and closures(2016 - 2021) Link
Description: This is a countrywide dataset of road construction and closure events, which covers 49 states of the US. Construction events in this dataset could be any roadwork, ranging from fixing pavements to substantial projects that could take months to finish. The data is collected from Jan 2016 to Dec 2021, using multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about 6.2 million construction and closure records in this dataset.

3. Methodology

Give us a rough idea on how you plan to use the datasets to answer these questions. (2 points)

- Data Collection:
  Download car accidents dataset from Kaggle, please check the link.
- Data Exploration:
  Use EDA to explore data to find main features and delete anomalies.
- Data Cleaning:
  Clean the data to deal with outliers, fill the blank values and make the distribution of data more reasonable.
- Data Integration: Might combine another two datasets about congestion link and road construction link.
- Data Analysis:
  Do you need to analyze data?
  Yes
  What analysis do you intend to do? (e.g., SQL, Statistics, Deep Learning)
  1. Analyze the correlation between factors (eg. weather) and the probability of car accidents.
  2. Use the existing dataset to figure out the most important factors.
  3. Divide data into training dataset and test dataset, and use machine learning algorithms to predict the probability of car accidents based on the previous important factors.

How to evaluate your analysis results? (e.g., evaluation metrics, confidence intervals, benchmark)

In the process of building predictive models, it's crucial to divide the dataset into training, validation, and testing sets. Initially, we employ the training set to build the model using advanced machine learning algorithms such as Random Forest. Subsequently, the validation set is used to fine-tune model parameters, assessing the model's performance by comparing predicted outcomes with actual results, ensuring the model's robustness and generalization ability.

- Data Product: What data product do you want to build? (e.g., visualization, an interactive web app, report, model)

  Design and create an User Interactive website which allows users to type in relevant data and get the prediction result.

4. Expected impact

Think about that once your project is complete, what impacts it can make. Pick up the greatest one and write it down. (1 point)

Once our project, focused on analyzing and predicting traffic accidents using the US Car Accidents dataset, is complete, the greatest impact it can make is significantly enhancing road safety. By identifying critical factors that contribute to traffic accidents, such as adverse weather conditions, high-risk locations, and times of day, our project can guide the implementation of targeted interventions. These interventions could range from improving road infrastructure in identified high-risk areas, optimizing traffic flow during peak accident times, to deploying more precise weather advisory systems for drivers. Ultimately, the project aims to reduce the number of traffic accidents by issuing accident warnings and indicating the risk level of potential accidents in given locations and weather conditions, thereby saving lives, reducing injuries, and minimizing traffic-related economic losses. This contribution to public safety and well-being would be the most significant impact of our life.

5. Potential challenges

Identify any anticipated obstacles and how you plan to address them. (1 point)

The datasets we have chosen are very large. The original data file exceeds 3 gigabytes, containing over 7.7 million rows of accident information and 46 columns of attributes (including time, location, weather, and accident details). Therefore, processing speed may be relatively slow, and analyzing the patterns among so many attributes presents a challenge. Additionally, the large amount of missing data in the dataset, which can negatively impact the results of data analysis and the performance of the accident prediction model.

To address these issues, we plan to:

- Utilize distributed computing frameworks (such as Apache Spark) to parallelize the processing of large-scale data, improving processing speed and efficiency.
- Employ interpolation, sampling, and other methods to fill in missing values.
- Perform data dimensionality reduction: Use dimensionality reduction techniques (such as principal component analysis) to reduce the dimensions of the data, thereby reducing computational burden and speeding up processing.
- Use incremental learning algorithms to gradually process and model large-scale data, rather than loading the entire dataset at once, thereby reducing the pressure on memory and computational resources.