DATASCI 451 Project Proposal

GROUP 21

November 2024

Group members names:

- **Xieyao Yin**
- **Rachel Rubanguka Hoops**
- **Jeremy Liu**
- **Casey Lee**

1 Research Question

What are the key factors contributing to traffic accidents in the United States, and how can we estimate accident probabilities?

2 Purpose of Data Analysis

2.1 What is the problem of interest?

The main objective is to identify and analyze factors contributing to traffic accidents and estimate accident probabilities based on weather, time of day, and road conditions. This analysis aims to improve traffic safety and inform interventions.

3 Objective

3.1 What is the final goal you want to achieve?

The final goal is to build a model that can predict the probability of accidents occurring based on environmental and temporal factors. This model aims to support data-driven traffic safety interventions.

4 Method

Use Bayesian Logistic Regression to analyze accident data, estimating the likelihood of accidents based on environmental and temporal factors.

5 Data

- **Is the data a simple random sample from some population?**

  No, the data is not a simple random sample. It consists of incidents recorded by traffic APIs, which may overrepresent areas with higher sensor coverage or more reporting agencies, leading to potential bias in coverage.

- **What is the potential defect in the data?**

  One potential issue is the non-uniform coverage across regions and possible gaps due to network connectivity. Additionally, some accident reports may be missing or inconsistently reported, and the data may overrepresent specific times, regions, or conditions (e.g., urban areas with dense sensor networks).

- **If you were able to collect the data, what would you have done differently?**

  Ideally, I would ensure consistent coverage across all regions and time periods, focusing on rural and less-monitored areas to balance the dataset. I would also verify data quality from all sources to minimize biases.

- **Are you going to account for the data collecting process in your analysis?**

  Yes, We plan to account for this by acknowledging potential biases due to data collection processes, possibly using weighting or stratification in the model to address over- or under-represented regions.

- **Describe simple facts about the data: size, data structure and source, etc.**

  The sampled dataset contains 500,000 records of traffic accidents. Each record includes details such as accident location, time, weather conditions, and road conditions. The data is in tabular format with variables representing categorical, numerical, and time-based information. It was collected from public traffic APIs used by various transportation and law enforcement entities.

- **If you are not going to use real data, explain how you are going to generate "realistic" synthetic data that represents your knowledge of the problem.**

  If We were to use synthetic data, We would generate it by simulating variables such as accident location, time of day, weather conditions, and road conditions based on known patterns in accident data. For example, We could increase accident probabilities under adverse weather conditions, during peak traffic hours, and in high-risk areas to mimic real-world trends. We would also base distributions for each variable on existing research or historical accident data to ensure the synthetic data reflects realistic conditions and probabilities.

5.1 Data Processing

6 Workflow, timeline, and responsibilities

6.1 Workflow

- **If you are working on a final project on your own, give a solid reason.**

  N/A

- **How are you going to divide the work? Do you have a group leader or convener? How often are you going to meet with each other?**

  We will do the coding, predictor selection, and analysis together. The report and presentation workload will be handled individually. We plan to meet once after the proposal and coding is finished, and will continue to communicate our progress online.

6.2 Tentative Timeline

Our team has divided the work to ensure that each member contributes to different aspects of the project. Here's the plan:

- **Group Leader/Convener:** One member(Rachel) will serve as the group leader to coordinate tasks and manage deadlines.
- **Meetings:** We plan to meet weekly to review progress, discuss challenges, and make any necessary adjustments to our plan.
- **Work Division:**
  - **Data Cleaning and Preprocessing:** One member will focus on cleaning the data and preparing it for analysis, ensuring consistent formats and handling missing values.
  - **Exploratory Data Analysis (EDA):** Another member will conduct EDA to summarize the data, identify patterns, and visualize key variables.
  - **Model Development:** One or two members will focus on developing the Bayesian Logistic Regression model, including feature selection and parameter tuning.
  - **Results Interpretation and Reporting:** The remaining members will interpret the results, prepare visualizations, and contribute to writing the final report.

For each group member (if you work on group projects): What is your role in the group project? What is your contribution so far? How do you plan to contribute further throughout the whole project?

Each member has a designated role in the project:

- **Data Cleaner:** Xieyao Yin - filtered the dataset to keep only relevant columns, processed categorical data transformation
- **EDA Specialist:** Jeremy Liu - will work on model and provide graphs to assess convergence and model accuracy
- **Model Developer(s):** Casey Lee - will work on Bayesian Logistic Regression model
- **Interpreter/Reporter:** Rachel Hoops - working on analysis and report

7 Draft a Plan for Data Analysis

7.1 What statistical model are you going to use for the data analysis?

We plan to use **Bayesian Logistic Regression** for our analysis. This model will allow us to estimate the probability of accidents occurring based on environmental factors (e.g., weather conditions) and temporal factors (e.g., time of day). Bayesian Logistic Regression is particularly useful here because it not only predicts accident probabilities but also provides insights into uncertainty, which is valuable for understanding risk under various conditions.

7.2 Have you done any preliminary analysis of the data?

We did initial inspection of the data which reveals the dataset structure. Data cleaning and key transformations were performed to drop irrelevant columns and handle missing values. Summary statistics and visualizations(e.g., severity distribution, accidents over time) are generated to better interpret the pattern.

7.3 What is your group's timeline for:

- **Propose and fit a preliminary model:**

  We aim to propose and fit an initial Bayesian Logistic Regression model by using the Metropolis MCMC algorithm to estimate β. This is necessary to complete the rest of the project and we aim to complete this by 11-19-24.

- **Examine initial results and adjust models if needed:**

  We will review the initial model results, evaluate performance, check the model with posterior predictive distribution, and adjust the model if necessary (e.g., refining feature selection and tuning model hyperparameters). This should be completed soon after the preliminary model by 11-21-24.

- **Prepare for a presentation:**

  After finalizing the model and completing our analysis, we will collaboratively prepare the presentation and add visualizations to communicate our findings. We aim to complete the presentation itself by 12-2-24.

- **Summarize and write the final report:**

  We will complete the final report together with the presentation, summarizing our analysis, model findings, and conclusions.

7.4 What are the (potential) difficulties with your data analysis?

Potential challenges include:

- **Handling Data Size:**

  The dataset is large. Managing memory usage while ensuring efficient processing may be challenging.

- **Addressing Biases:**

  Since the data collection process may have regional and temporal biases, we need to consider

how these might affect our results and explain these effects in our report.

- **Model Convergence:**

  Bayesian models can be computationally intensive. Achieving convergence in the logistic regression model might be difficult, especially if the data is highly imbalanced or if certain variables lack sufficient variation.

- **Interpretation of Results:**

  Bayesian logistic regression provides probability estimates with credible intervals. Clearly interpreting and communicating these results in a way that is understandable to all audiences could be challenging.

8 Acknowledgments

A sampled version includes 500,000 accidents extracted from the original dataset for easier handling and analysis. The dataset can be accessed at the following link:

https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

Papers citations from the author:

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. ”Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights.” In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.