Applied Analytics Project

Analyzing US Accident Data to Predict High-Risk Areas and Times in Massachusetts

Major: Applied Analytics

Name: Gefan Wang, Chenhe Shi, Tianchen Liu

Date: 01.26.2025

Project Overview:

The primary goal of this project is to analyze the Kaggle US Accident dataset to uncover key factors influencing the number of accidents in specific areas of Massachusetts. By focusing on variables such as severity, weather, geographic location, infrastructure, we aim to develop a predictive model that identifies areas and hours with a high likelihood of accidents. This model will be instrumental for stakeholders, including businesses such as delivery companies and government agencies, in optimizing their operations, minimizing disruptions, and reducing associated costs.

Hypothesis:

Certain factors such as adverse weather conditions, specific geographical features, and infrastructures have a significant impact on the frequency and severity of accidents. By identifying these key factors, it is possible to reduce the number of accidents and improve operational efficiency for businesses and government entities.

Economic Value Calculation:

Accidents disrupt transportation-dependent businesses, such as police departments and delivery services. These disruptions lead to increased costs, including vehicle damage, delays, resource utilization, and potential reputational harm.

Base Assumptions:

Over six years, the dataset records 62,000 accidents in Massachusetts. There are 1.9 million registered cars in Massachusetts. There are 1 million people who rely primarily on transportation. Current accident-related disruptions occur in 10% of cases (positive hits). Positive hits refer to accurately predicting an accident-prone area or time to mitigate potential disruptions, which lets say it is 30% now. Negative impacts (false negatives) refer to instances where the model fails to predict an accident in a high-risk scenario, leading to unanticipated disruptions, occurring in 10% of cases.

Assume after we improve the model, there is 60% accuracy in predicting positive hits and still 10% chance of false negatives (Since more signs and police in the area will only raise people's alert when driving). Using the weighted calculation: 60% × 30% + 10% × (-10%) = 17% net impact improvement. Improved prediction rate: 30% × (1 + 17%) = 35.1%. Total businesses positively impacted: 351,000 (35.1% of 1 million MAU). Assume the average cost per accident for businesses is $5,000 (including direct and indirect costs). Total cost savings = (35.1/62,000) × $5,000 / 6 = $18.14 million annually.

Project Plan:

Week 1-2: Data Exploration and Cleaning

- Load and filter the dataset for Massachusetts.
- Handle missing values and remove irrelevant columns.
- Perform initial exploratory data analysis .

Week 3-4: Feature Engineering

- Create new features (accident frequency by location, weather categories).
- Extract insights from the "Description" column using NLP techniques, such as identifying keywords or phrases related to accident causes.
- Normalize and encode categorical variables.
- Conduct correlation analysis to identify key predictors.

Week 5-6: Model Selection and Preparation

- Split the dataset into training and testing sets.
- Select appropriate supervised learning models (decision trees, random forests, etc).
- Define metrics for evaluation (accuracy, precision, recall).

Week 7-9: Model Training and Evaluation

- Train selected models on the training dataset.
- Evaluate performance on the test dataset.
- Perform hyperparameter tuning for optimal results.

Week 10: Insight Generation

- Summarize key findings from models.
- Identify actionable recommendations for policymakers and stakeholders.

Week 11-12: Report and Visualization

- Prepare visualizations (maps, graphs, etc) to convey insights.
- Draft a comprehensive report documenting methodology, findings, and recommendations.

Week 13-14: Presentation and Delivery

- Prepare presentation slides.
- Submit the report, visualizations, and code to GitHub.
- Present findings to stakeholders.

About the dataset:

The dataset is sourced from Kaggle and contains information around 62,000 accidents in Massachusetts from February 2016 to March 2023. It includes data on geographic locations, weather conditions, and infrastructure features (presence of bumps, crossings, or amenities). Key attributes of the dataset:

- Geographic Information: Latitude, longitude, city, state, and zip code.
- Weather Data: Temperature, humidity, visibility, and precipitation.
- Infrastructure: Indicators for bumps, crossings, and amenities.
- Time Data: Start and end time of incidents.
- Description: Text descriptions of the accidents, which may contain additional valuable details.

Type of Modeling:

The project will use supervised learning to predict the severity of accidents based on weather, infrastructure, and location data. This is a classification problem with a multi-class target variable (Severity levels 1 to 4).

If exploratory analysis reveals clusters of high-risk areas or conditions, unsupervised learning might be used to identify patterns.

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