Applied Analytics Project

Pocket Radar: A Data-Driven Approach to Pinpoint Accident Hotspots in Massachusetts

Final Report

Major: Applied Analytics

Name: Gefan Wang, Chenhe Shi, Tianchen Liu

Date: 05.04.2025

Project Overview

As students and drivers in Massachusetts, we deeply care about road safety and its impacts on our community. Road accidents in Massachusetts cause significant disruptions, financial losses, and safety concerns for local communities, businesses, and public agencies. To make these insights actionable, we propose an **accident simulation system** that provides real-time severity heatmaps for drivers, police officers, and government officials. This enables proactive interventions such as targeted law enforcement deployment, improved route planning for commercial and public transportation, and infrastructure. Our findings offer a data-driven framework to predict accident severity, enhance roadway safety, and minimize economic losses for businesses and municipalities.

Model Development Life Cycle (MDLC)

The Model Development Life Cycle consists the following major steps:

1. **Business Understanding:** Define objectives to predict accident severity and mitigate economic and safety impacts.
2. **Data Collection:** Consists over 60,000 accident records from Massachusetts (2016-2023).
3. **Data Preprocessing:** Managed missing values, performed feature engineering including temporal and environmental variables, and categorized accident severity.
4. **Exploratory Data Analysis (EDA):** Identified key risk factors like adverse weather conditions and peak traffic times.
5. **Model Selection:** Evaluated multiple algorithms (KNN, Decision Trees, Random Forest, XGBoost, Naive Bayes, PCA + Logistic Regression) and addressed class imbalance through stratified sampling and SMOTE.
6. **Model Training and Validation:** Conducted extensive hyperparameter tuning and validated using accuracy, precision, recall, and F1-score metrics.
7. **Deployment:** Created an interactive website delivering real-time accident severity heatmaps for practical interventions.
8. **Evaluation and Monitoring:** Monitored performance metrics regularly to ensure continued accuracy and effectiveness.

Data Description and Preparation

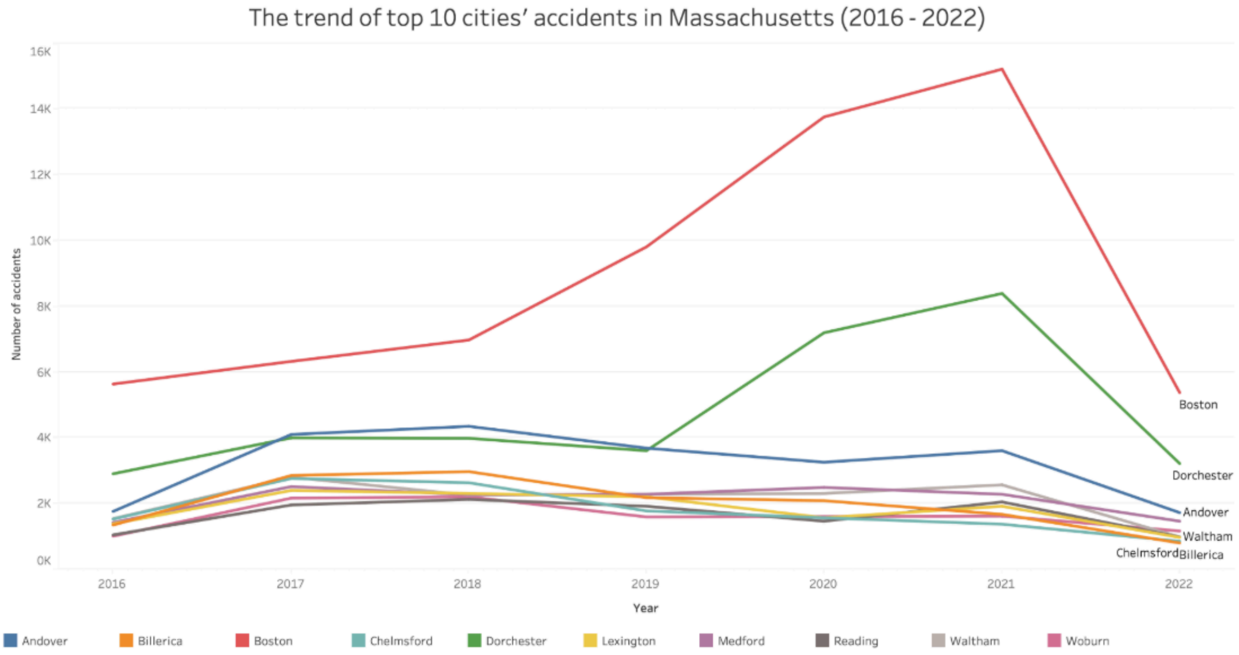
We utilized a comprehensive dataset containing over **60,000 accident** records from 2016 to early 2023.

- Columns with substantial missing data were removed.
- Missing numerical values were filled using median imputation, while categorical data used mode imputation.
- Temporal features (hour, day of the week, month), environmental conditions (temperature, visibility, precipitation), and infrastructural features were engineered to enhance predictive power.
- Accident severity was categorized into four classes ranging from minor crashes without injuries (Class 1) to major crashes with fatalities (Class 4).

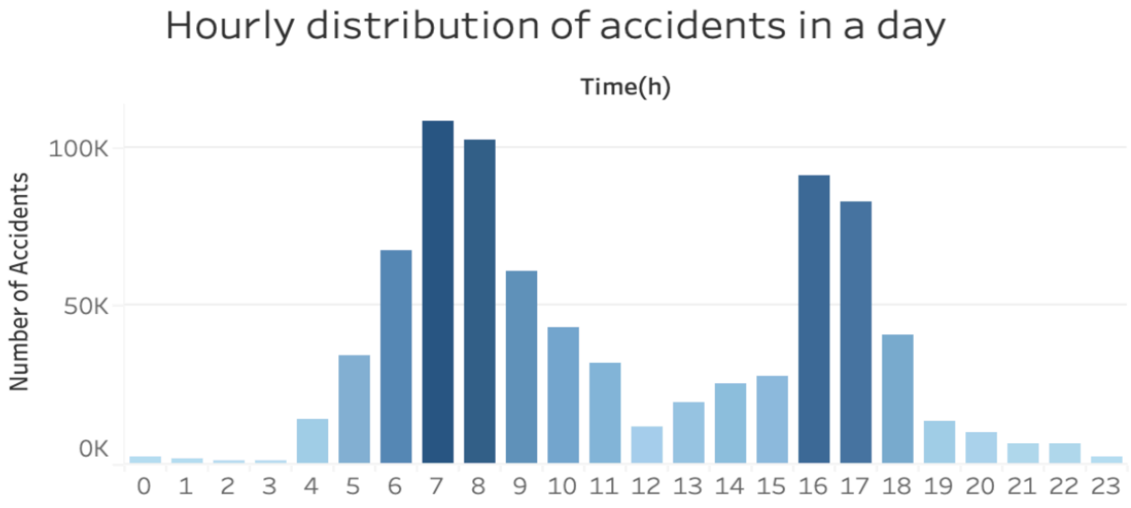
Exploratory Data Analysis (EDA)

EDA provided critical insights into the data, highlighting accident trends:

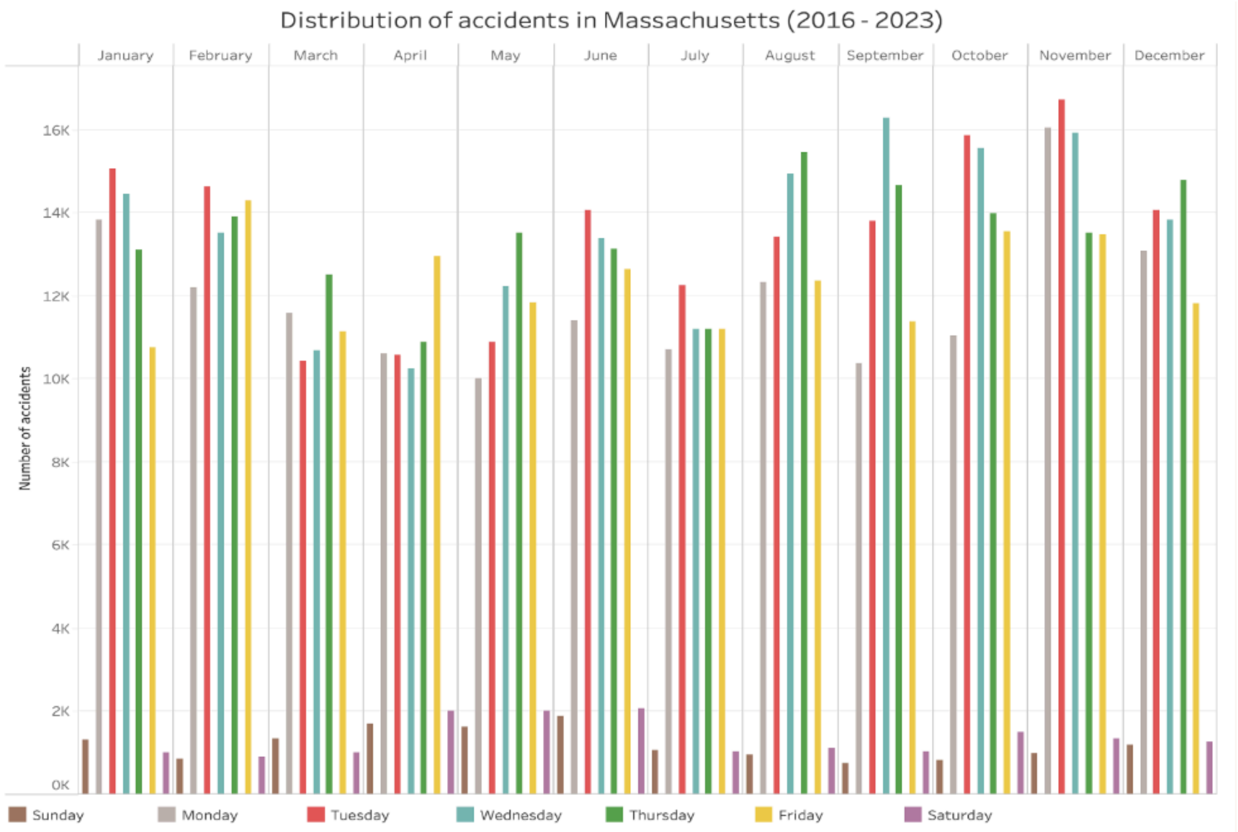
- Geographical Analysis: Most accidents occurred in major urban centers, with notable peaks in cities such as Boston.



- Hourly Trends: Accident frequencies were highest during morning (7-9 AM) and evening (4-6 PM) rush hours.



- Seasonal Patterns: Accident rates were slightly higher during winter months compared to other seasons, likely due to adverse weather conditions.



Model Development and Evaluation

Initially, our model comparisons included PCA + Logistic Regression, Decision Trees, Random Forest, XGBoost, KNN, and Naive Bayes. After further iterations, including extensive hyperparameter tuning and more robust validation, **XGBoost** emerged as the optimal model choice due to its higher accuracy and robustness against data imbalance.

The final evaluation metrics, including accuracy, precision, recall, and F1-scores, confirmed that XGBoost consistently outperformed other models, demonstrating improved accuracy particularly in predicting severe accidents (Classes 3 & 4).

Results

The optimized XGBoost model delivered superior performance with an accuracy significantly surpassing the initial PCA + Logistic Regression benchmark (initially at 72%). Detailed model metrics indicated substantial improvements in predicting rare, severe accidents, thus meeting our goal of effectively identifying critical accident hotspots.

Implementation and Impact

We developed an interactive website providing real-time severity heatmaps, enabling users—drivers, law enforcement agencies, city planners—to visualize accident risk dynamically. Our implementation facilitates immediate practical interventions, such as improved law enforcement deployment and route adjustments for logistics services.

$$Total\,Saving\ =\ (\sum_{i=2}^{4}\frac{Total\,Cases_i}{6} * Cost\,Per\,Case_i) * 0.3$$

By using this equation above, we projected approximately $295 million in annual savings through reduced accident-related costs, benefiting both governmental and private sectors.

Conclusion

Through rigorous data-driven techniques, we successfully developed a highly predictive model that significantly advances road safety management. This project underscores the critical role of machine learning in addressing public safety challenges, providing actionable insights, and substantially reducing economic burdens.

Future Work

Future efforts will focus on integrating the model directly into widely-used navigation platforms such as Google Maps or Apple Maps. This integration will enable real-time alerts about potential accident hotspots, enhancing proactive decision-making and significantly improving overall road safety.

Work Cited