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

Gefan Wang, Tianchen Liu, Chenhe Shi
Woods College of Advancing Studies, Boston College

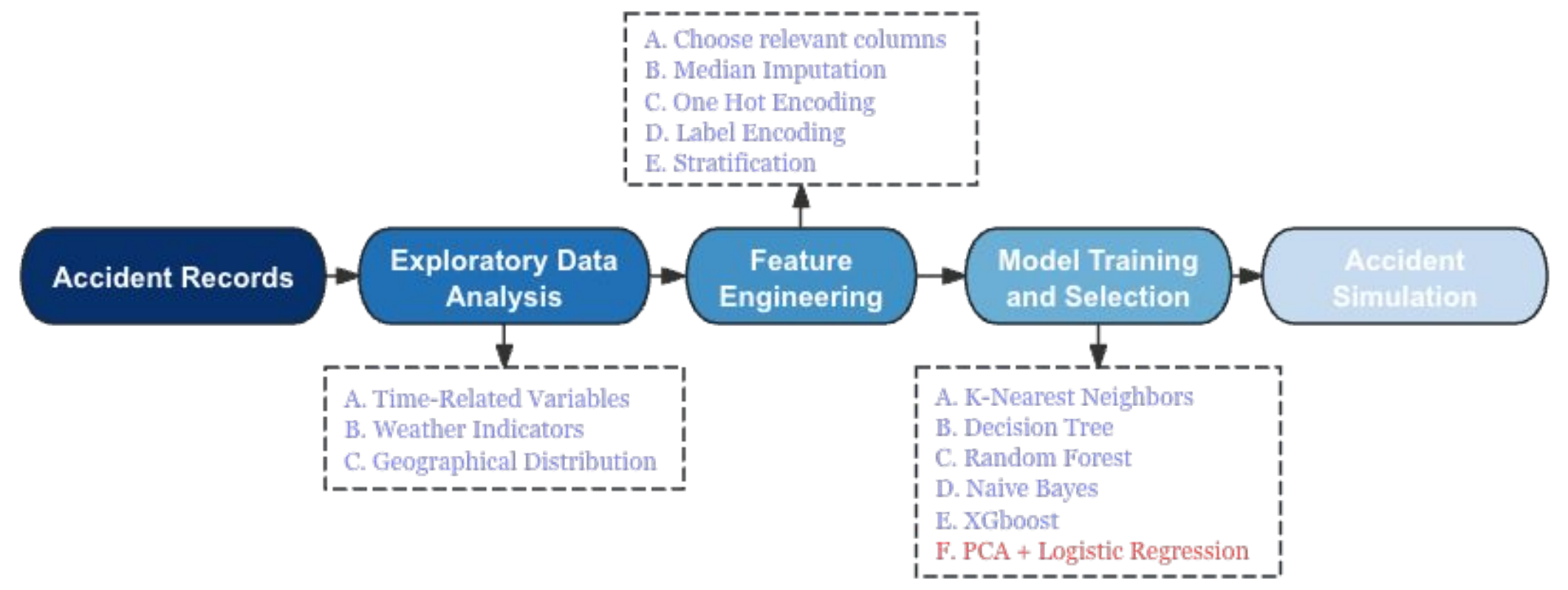
Abstract

Road accidents in Massachusetts cause significant disruptions, financial losses, and safety concerns for local communities, businesses, and public agencies. **This study leverages a dataset of over 60,000 accidents, spanning the period from 2016 to 2023, to examine how environmental conditions, infrastructural features, and time-related factors (e.g., hour of day, day of week, rush-hour indicators) influence crash severity.**

Through our exploratory data analysis (EDA), we discover trends hidden in poor weather conditions—such as low visibility and heavy precipitation—leading to higher-severity collisions, as well as the role of traffic junctions and urban density in accident frequency.

To make these insights actionable, we propose a **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.

Project Pipeline & Research Question



Our pipeline processes accident records through **exploratory data analysis (EDA)**, extracting **time-related, weather, and geographical features**. We apply **feature engineering** techniques like encoding, imputation, and stratification. Various **machine learning models** are trained and evaluated, with **PCALogReg** outperforming others.

Core Objective: Identify the driving factors of crash severity in Massachusetts—time variables, spanning weather, and infrastructures—to build a predictive model that flags areas with high-severity accidents that are likely to happen.

Data Source & Preparation

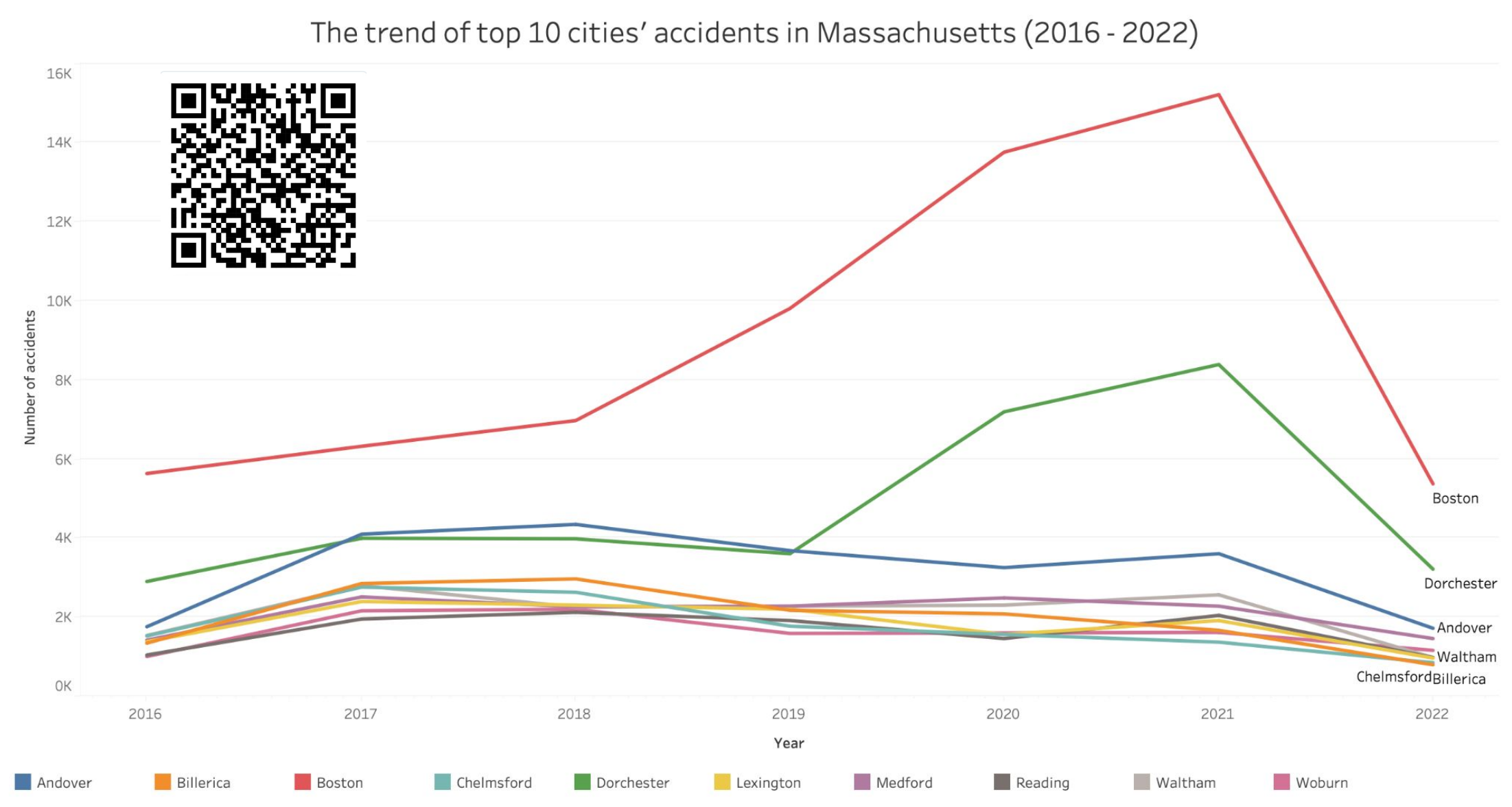
Dataset: Over 60,000 accident records, drawn from the US_Accidents_MA.csv, covering Massachusetts from 2016 through early 2023.

Cleaning & Filtering:

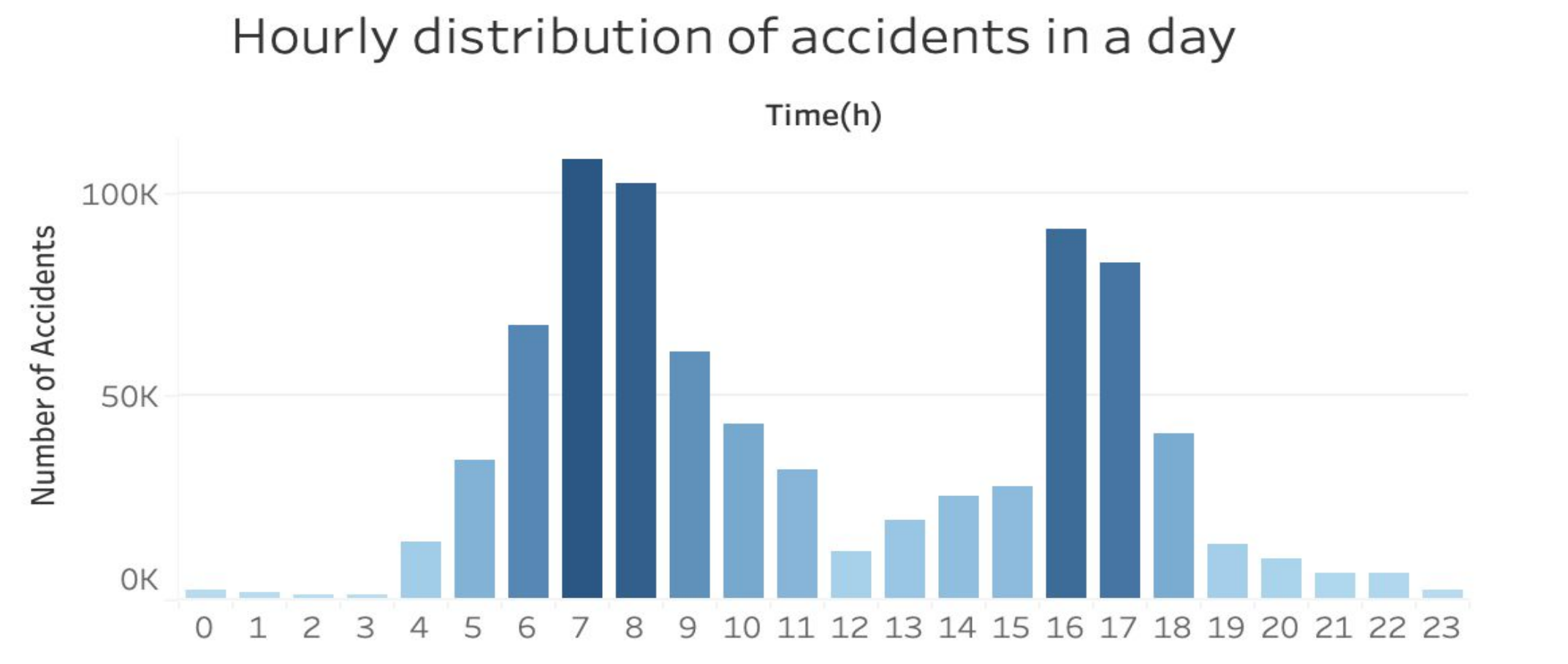
- Dropped columns with excessive missingness (e.g., End_Lat, End_Lng if over 50% missing).
- Converted date/time columns (Start_Time, End_Time) into proper datetime formats for feature engineering.
- Imputed missing numerical variables (Temperature(F), Precipitation(in), Wind_Speed(mph), etc.) using median values and categorical variables (Weather_Condition, Wind_Direction) using mode.
- Defined accidents severity from 1-4 where 1 means minor crashes without injury, 2 means minor crashes with non-fatal injury, 3 means major crashes with non fatal injury, 4 means major crashes with fatal injury. Minor accidents does not cause significant traffic shutdown while major accidents causes traffic paralysis.

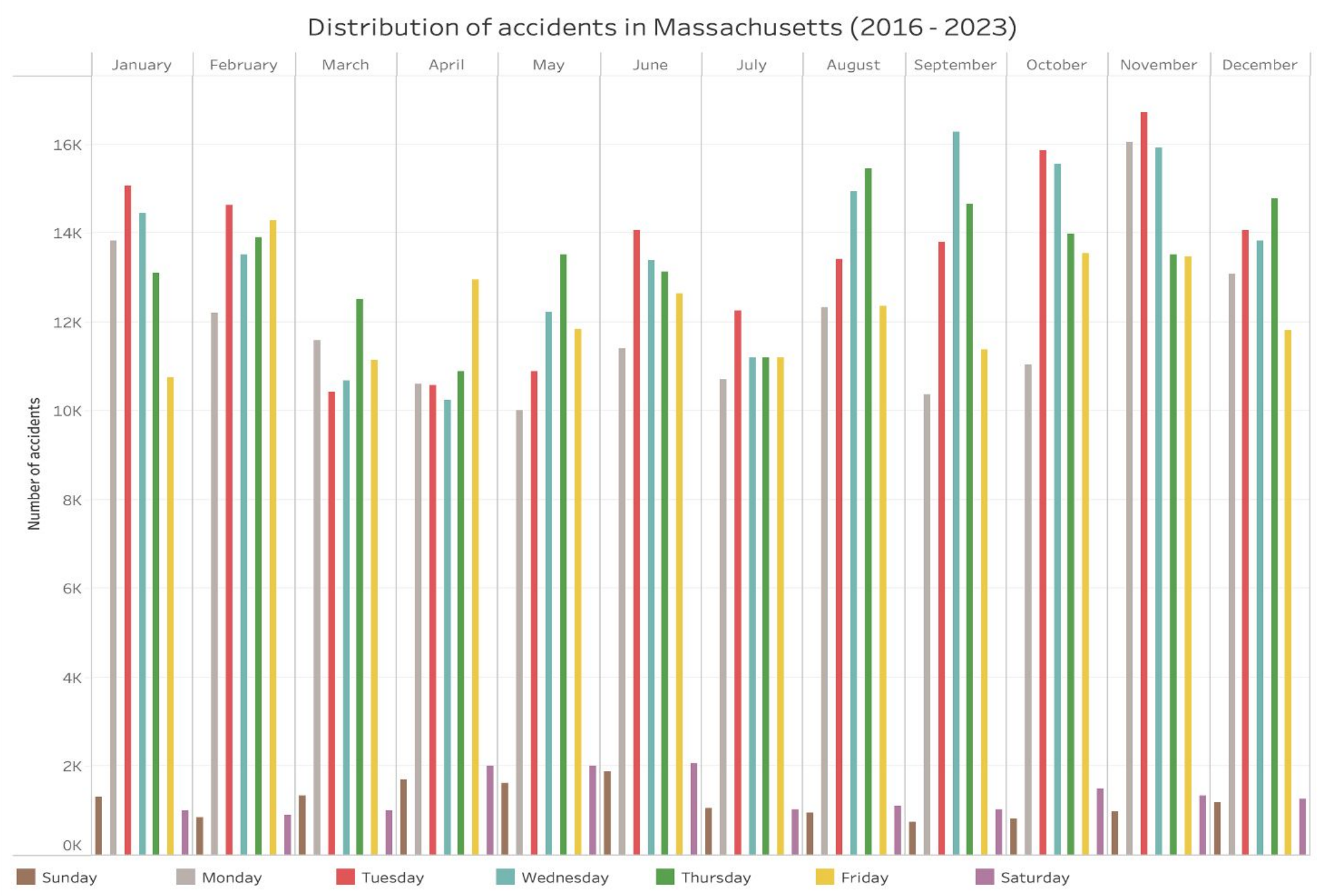
Exploratory Data Analysis (EDA)

This map illustrates the number of accidents happened in major cities in the state of Massachusetts from 2016 to 2022. Scan the QR code to see the interactive map for accidents severity score.

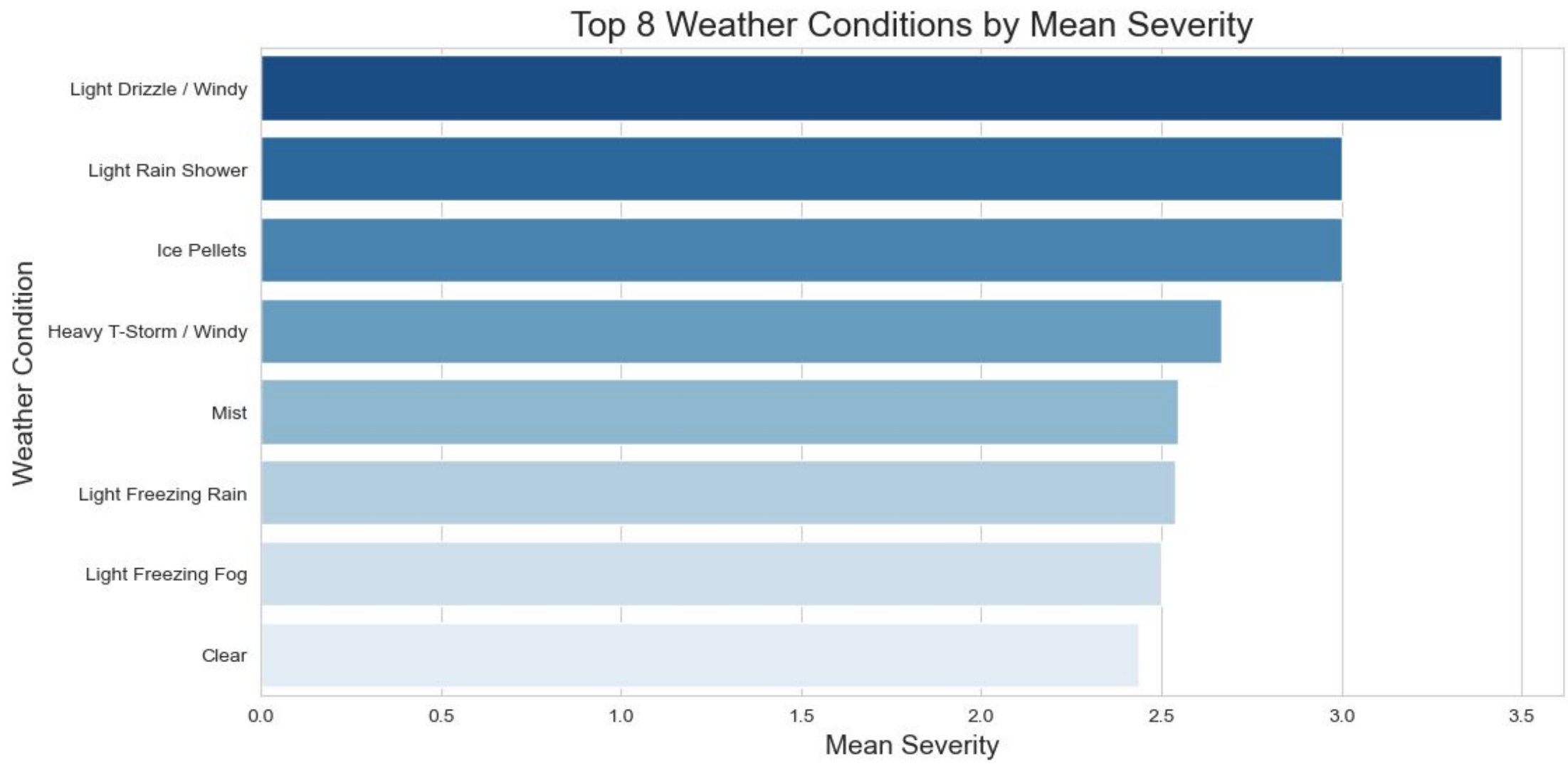


- **Time-related Variables**: Days of weeks, months of years, hours of a day. Our data suggests that most accidents occurs in **weekdays rush hours** than weekends with slightly more accidents happening in **winter** compared with other seasons of a year.





- **Weather Indicators**: Categorical grouping of weather conditions (rain, snow, fog). Our data indicates that accidents with highest severity happened in a **light rain and windy** weather condition. Drivers should be particularly cautious in poor weather conditions (light drizzle, mist) but it still pose significant risk.



Feature Engineering

- We engineered **temporal, environmental, infrastructural, and distance-based** features to improve model performance. These features allow the model to learn how **time of day, weather, and road conditions** impact accident severity.

Feature Name	Description	Type
Hour	Hour of the accident (0-23)	Numeric
Is_Weekend	1 if Sat/Sun, 0 otherwise	Binary
Weather_Condition	Grouped into Clear, Rain, Snow, Fog	Categorical
Visibility_Category	High, Medium, Low	Categorical
Junction	1 if accident at an intersection	Binary

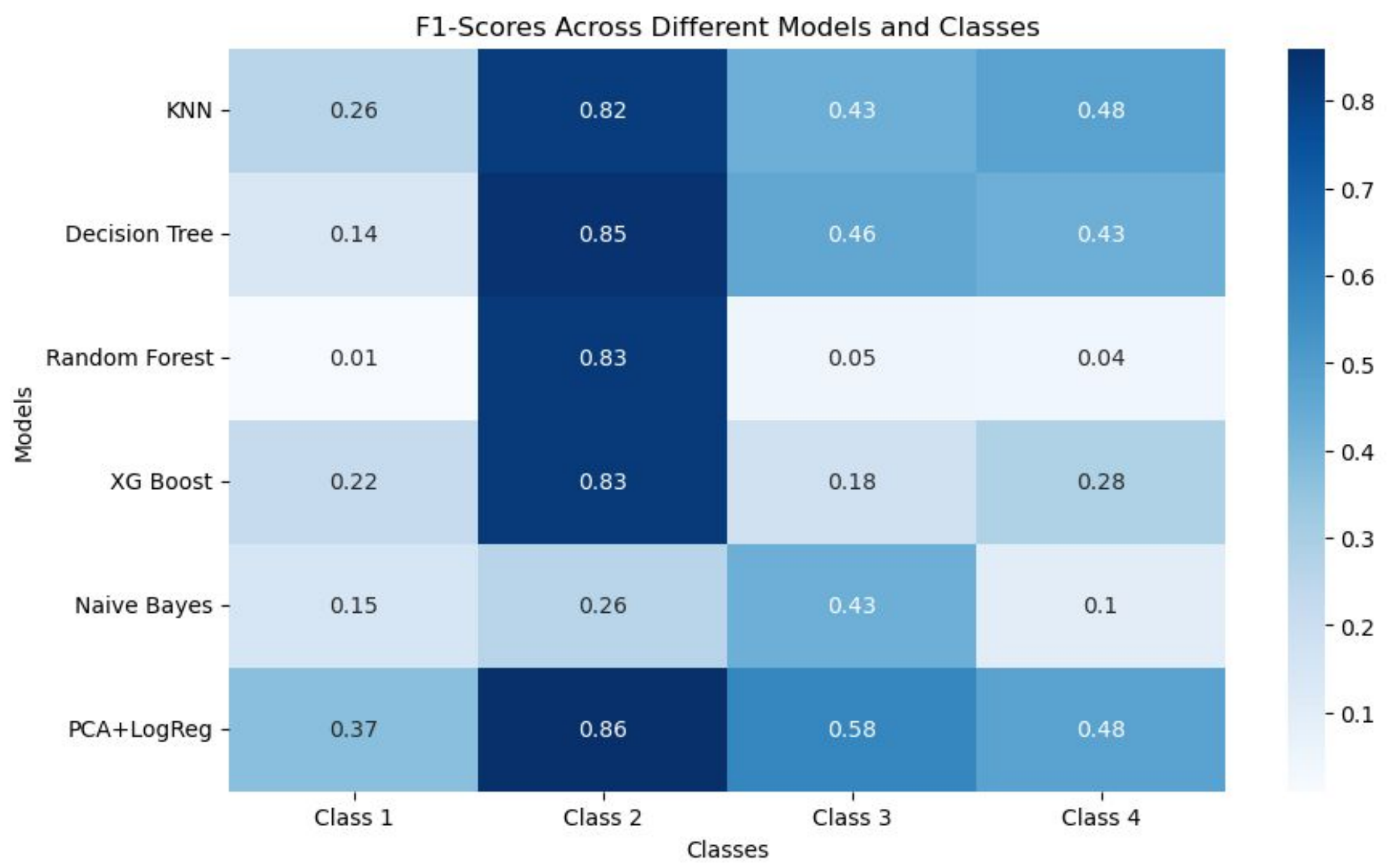
Results

Modeling & Evaluation

- **Severity Classes**: We classified accidents into four levels of severity: 1)Minor crash with no injuries, 2) Minor crash with non-fatal injuries, 3)Major crash with serious but non-fatal injuries, 4) Major crash with fatal injuries. Classes 3 & 4 form only about 10% of the dataset combined, requiring special handling.
- **Model Selection**: We compared KNN, Decision Trees, Random Forest, XGBoost, Naive Bayes, and PCA + Logistic Regression.
- **Imbalance Mitigation**: Due to fewer Class 3 & 4 samples, we applied stratified sampling and SMOTE oversampling to improve detection of severe accidents.
- **Metrics**: We measured **Accuracy**, **Precision**, **Recall**, and **F1-score** per class, to ensure high sensitivity to minority classes.

Modeling Performance

- Among these methods, **PCA+LogReg** achieved the highest validation accuracy (0.72), outperforming other models due to its ability to handle structured data, capture nonlinear relationships, and effectively weigh multiple risk factors.



- **Winning Model on Validation Set: PCA+LogReg**

- We also construct this website for government officials, city planners, police officers and drivers to input the real-time variables to get a **interactive map** of Massachusetts with predictions on real-time severity of accidents. Please **scan the following QR code** and try our website. Feel free to save this website as it could help you next time when you drive.

Hour: 0-23(hour of a day)
Month: 1-12(of a year)
Week: 1-7(Mon-Sun)
Weather Condition: Rain, Clear, Cloudy, Snow
Temperature: in Fahrenheit
Humidity: in percent
Pressure: in inch
Visibility: in mile
Wind Speed: in miles per hour
Wind Condition: wind direction

Discussion

Annual savings on costs:
Assume that 30% of accidents are reduced by enabling early intervention through law enforcement and infrastructure improvements, our model could help prevent or reduce accident severity, leading to **~$295 million in annual savings**.
Government Savings (40%): ~$118 million per year
Covers emergency response, public healthcare costs, road safety infrastructure, and law enforcement interventions.
Private Sector Savings (60%): ~$177 million per year
Includes insurance payouts, vehicle repairs, lost productivity, and logistics disruptions for businesses and individuals.

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

Where ***i*** is the severity level

Impacts of our models:
For Drivers & Public Safety: A QR code-based system delivers real-time severity heatmaps, helping drivers make informed decisions on high-risk routes. Additionally, by prioritizing emergency response readiness, this model could help reduce mortality and injury rates by enabling faster dispatch in high-severity accident zones.
For Policymakers & Law Enforcement: Insights from our model can support speed adjustments, improved signage, and optimized law enforcement deployment, ensuring safer road conditions.
For Businesses & Logistics: Ride-sharing and delivery services can integrate severity predictions into route planning, reducing accident-related delays and operational costs. We estimate a 17% improvement in identifying accident-prone areas, potentially saving up to $18.14 million annually in direct and indirect costs.

Conclusion & Future Work

Conclusion
Our study demonstrates the effectiveness of machine learning in predicting accident severity using time, environmental, and infrastructural features. **PCA+LogReg** achieved the best performance, highlighting key risk factors. Feature engineering, including time-based and weather-related variables, significantly improved predictive performance. Addressing class imbalance through stratified sampling and oversampling techniques enhanced severe accident detection.

Future Work
To enhance real-time accident prevention, we aim to integrate our model into Google or Apple Maps, allowing for dynamic risk alerts. When a driver approaches a high-severity area predicted by our model, the map could provide a real-time warning, indicating elevated accident risk based on current weather, traffic, and infrastructure conditions. This integration would increase driver awareness, improve proactive decision-making, and ultimately enhance road safety.

References

[1] Data source from: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents/data

[2] Wikipedia contributors. (n.d.). Traffic collision. Wikipedia, The Free Encyclopedia. Retrieved March 6, 2025

[3]Mass.gov. 2019. "Hwy-CrashCosts_2019_Dollars.pdf | Mass.gov." Mass.gov. 2019. https://www.mass.gov/doc/crash-costs-methodology.

[4]GeeksForGeeks. 2018. "Principal Component Analysis(PCA)." GeeksforGeeks. July 7, 2018. https://www.geeksforgeeks.org/principal-component-analysis-pca/.

[5]NHTSA. 2023. "NHTSA: Traffic Crashes Cost America $340 Billion in 2019 | NHTSA." Www.nhtsa.gov. January 10, 2023. https://www.nhtsa.gov/press-releases/traffic-crashes-cost-america-billions-2019.

BOSTON COLLEGE
Woods College of Advancing Studies