Traffic Collision Analysis: A Machine Learning Approach

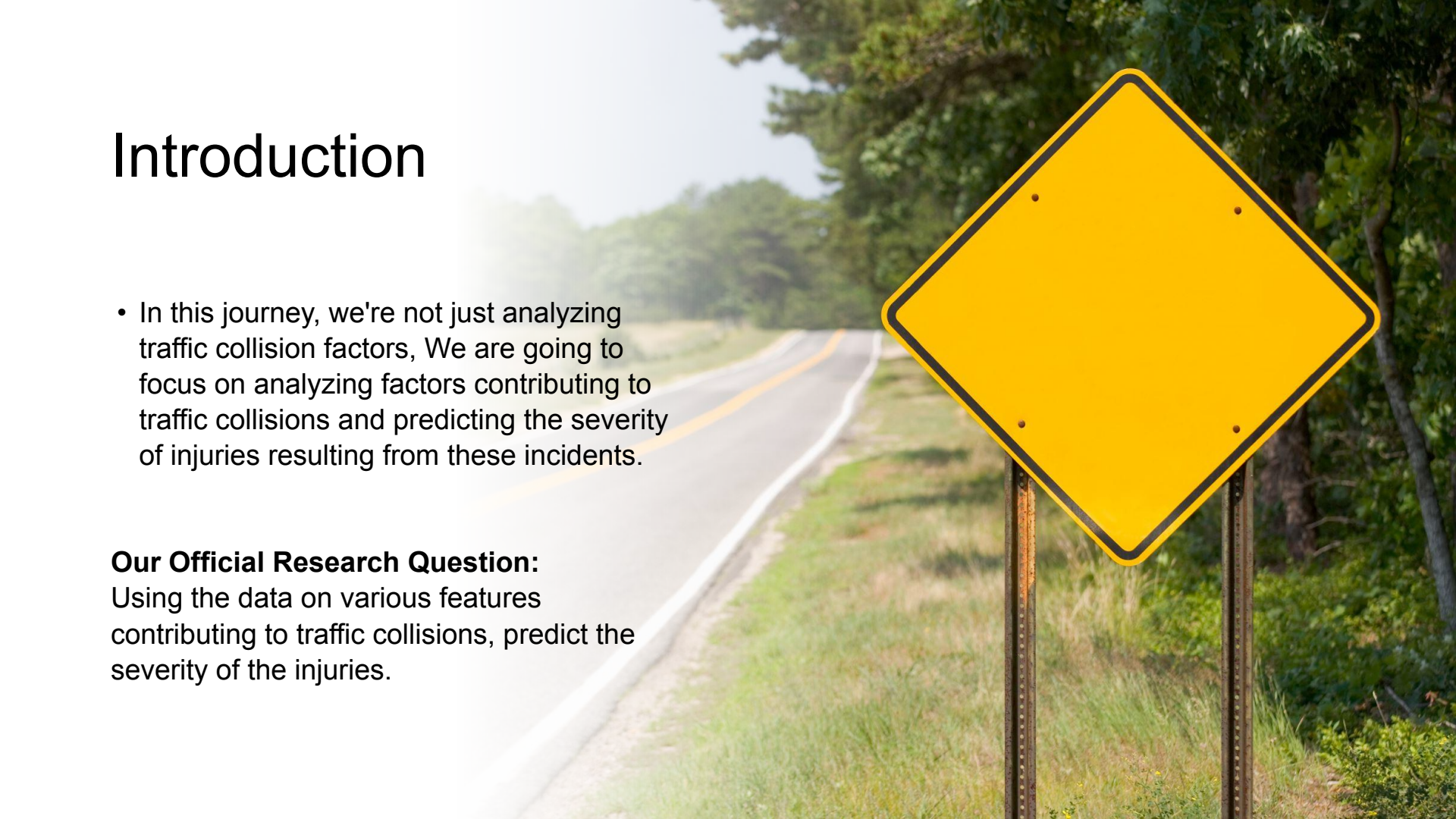
Diving Deep into Traffic Safety with Data and Humor

Presenters: Botian Zheng, Kenneth Choi, Zemin Chen, Shiyu Li, Demin Chen

Introduction

- In this journey, we're not just analyzing traffic collision factors, We are going to focus on analyzing factors contributing to traffic collisions and predicting the severity of injuries resulting from these incidents.

Our Official Research Question:
Using the data on various features contributing to traffic collisions, predict the severity of the injuries.



Context and Motivation

- Traffic collisions are one of society's major public safety issues, often resulting in injuries and fatalities of varying degrees and sizes and significant economic losses.
- Motivation: Improve the response speed for emergency department and decrease the risk of future traffic collisions.
- Increased awareness of collision risks and make safer driving practices

Stakeholders

Key players:

- Traffic safety authorities
- Hospitals
- Insurance companies

Their efficiency is our top priority, just like making sure you reach your favorite destination on time!

Dataset Part I

Data Origin Source:

- Data collected from the Statewide Integrated Traffic Records System (SWITRS), provided by TIMS at Berkeley (source).

Direct Source:

- Alex integrates those data（source）

Descriptive Analysis:

- Involves victim data with attributes like age, injury severity, safety equipment usage, etc.

Dataset Part II

Feature Engineering and Filtering:

- Conversion of categorical data into numerical formats, selection of relevant features impacting injury outcomes. Random selection of 10,000 records from the original dataset for training.

Data Categories:

- 'Killed or Severely Injured', 'No Injury', 'Other Visible Injury'.

Methods

- Machine Learning Algorithms:
  - Use of Logistic Regression, SVC, KNeighborsClassifier, LinearSVC, DecisionTreeClassifier, and RandomForestClassifier.
  - Application of data standardization techniques to enhance model performance.

- Model Selection Rationale:
  - Each model offers unique strengths in pattern recognition and prediction accuracy

Results and Conclusions

- Model Performance:
  - Evaluation of models based on accuracy, with RandomForestClassifier and DecisionTreeClassifier showing superior performance.

```
Logistic Regression Precision: 0.6821236924483935
SVC Precision: 0.6871673922335827
KNeighbors Precision: 0.7332548738348884
Linear SVC Precision: 0.6758891760956661
Decision Tree Precision: 0.7361879238227411
Random Forest Precision: 0.7536459614042762
```

- Key Findings:
  - Identification of significant predictors of injury severity in traffic accidents.
  - Insights into the relationships between various factors and collision outcomes.

	precision	recall	f1-score	support
0	0.41	0.08	0.13	195
1	0.91	0.97	0.94	279
2	0.79	0.94	0.86	743
accuracy			0.81	1217
macro avg	0.70	0.66	0.64	1217
weighted avg	0.76	0.81	0.76	1217

Results and Conclusions

- Readiness:
  - Assessment of the models' readiness for real-world application, including the need for broader dataset validation and model refinement.
  -
- Real World Applications:
  - Enhancing traffic safety measures
  - informing hospital emergency preparedness
  - accurate risk assessment for insurance

Goal: Create a Predictive Model on the SWITRS Dataset

Recap

	precision	recall	f1-score	support
0	0.41	0.08	0.13	195
1	0.91	0.97	0.94	279
2	0.79	0.94	0.86	743
accuracy			0.81	1217
macro avg	0.70	0.66	0.64	1217
weighted avg	0.76	0.81	0.76	1217

Dataset:

- SWITRS by TIMS on traffic accident data
- Convert Categorical Features and Random Sampling of Dataset

Methods:

- Standardized data
- Used 6 methods ensembled together

Results:

- Random Forests ≈ 0.734
- Decisions Trees ≈ 0.736
- Good at finding some and no injuries
- Hard time finding fatal/severe injuries
  - 3 outputs vs 2 outputs

Future Directions

- Expanding into new datasets with diverse new features to better our model is essential
- Figuring out new methods of categorical predictions is needed
    - Predict 2+ variables consistently
- Prediction can expand to new stakeholders
- Create new products out of prediction data