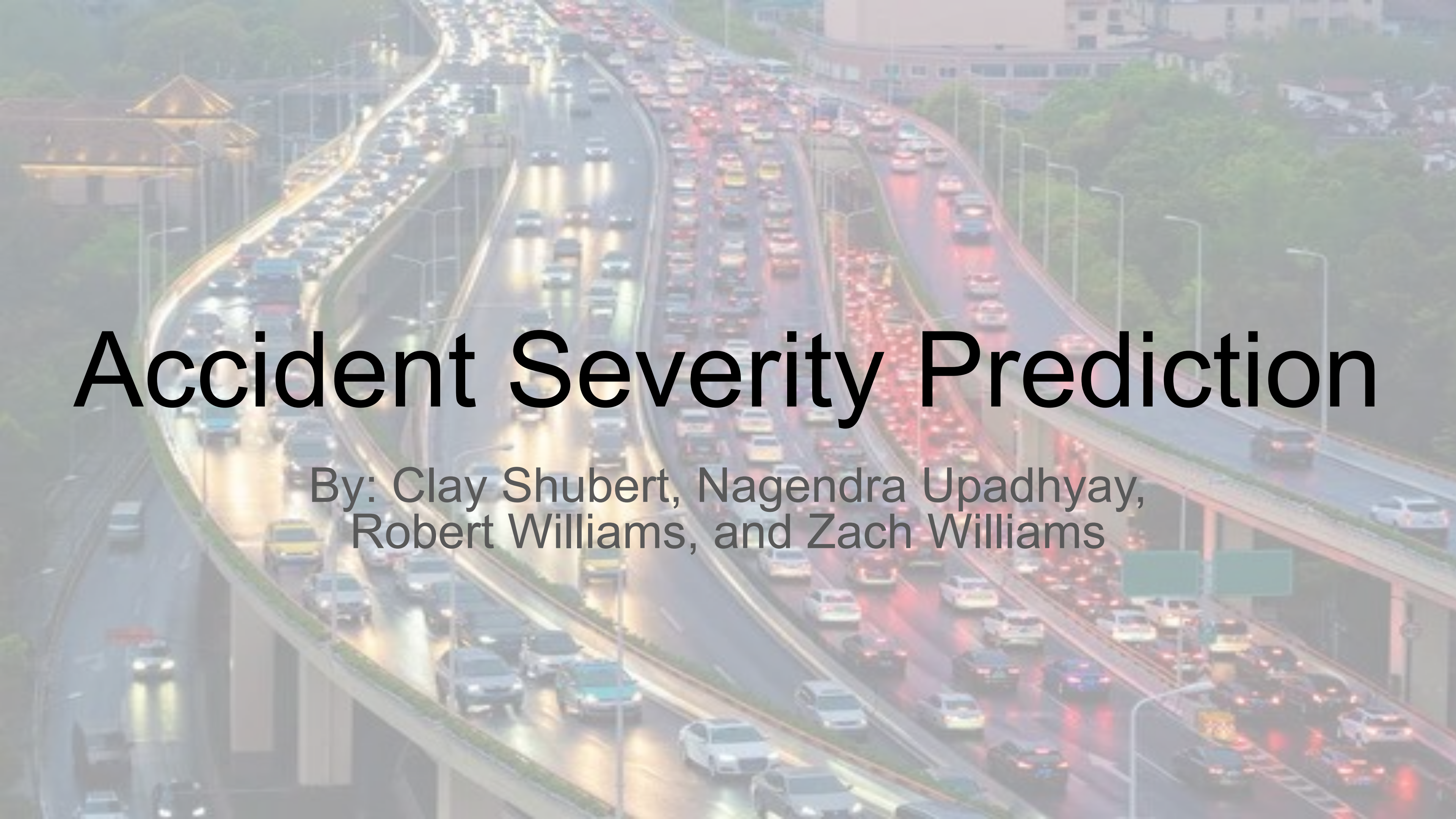
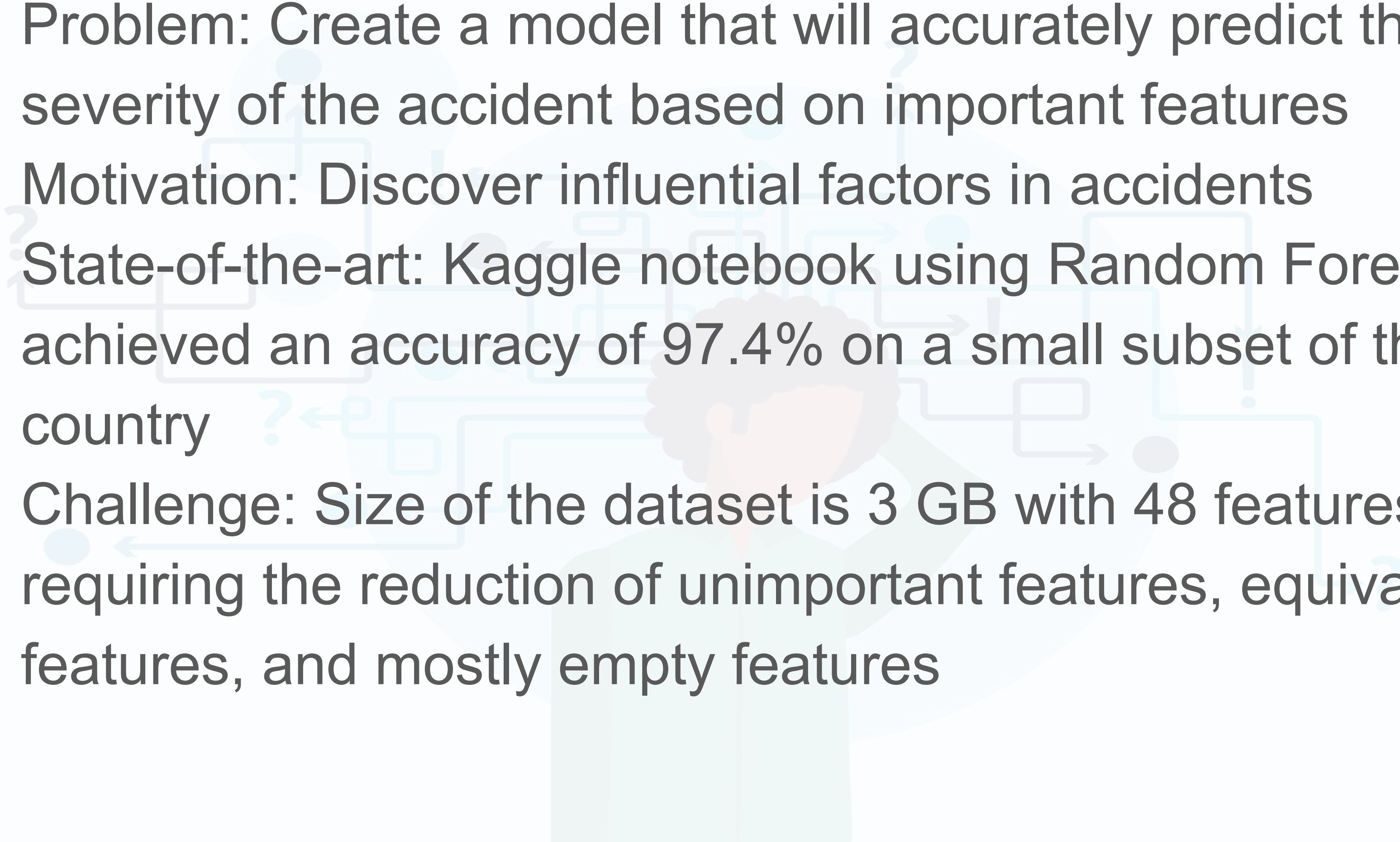


Accident Severity Prediction

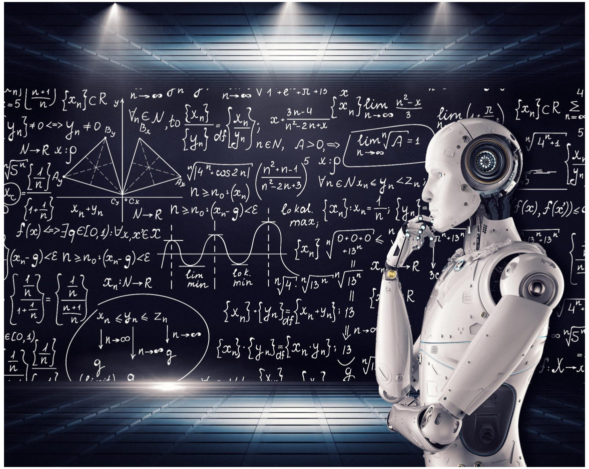
By: Clay Shubert, Nagendra Upadhyay,
Robert Williams, and Zach Williams

- Problem: Create a model that will accurately predict the severity of the accident based on important features
- Motivation: Discover influential factors in accidents
- State-of-the-art: Kaggle notebook using Random Forest achieved an accuracy of 97.4% on a small subset of the country
- Challenge: Size of the dataset is 3 GB with 48 features requiring the reduction of unimportant features, equivalent features, and mostly empty features

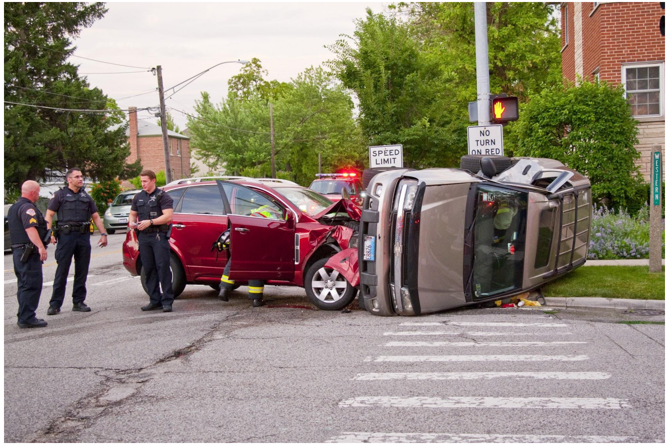


Problem

- Our group wants to understand environmental and road factors that have an impact on the severity of traffic caused by an accident
- We are also interested in the correlation between various features in the dataset
- The analysis of car accidents is a very important topic as it can help us to understand what needs to be done to prevent accidents, or decrease their overall impact on traffic



Motivation



- Car accidents are one of the largest causes of death globally
- A wide range of factors influence car accidents and their impact on traffic, so it is a complex problem to try and analyze
- Effectively predicting influential factors can give insight to community leaders on ways to reduce the likelihood of a car accident or the impact on traffic caused by it

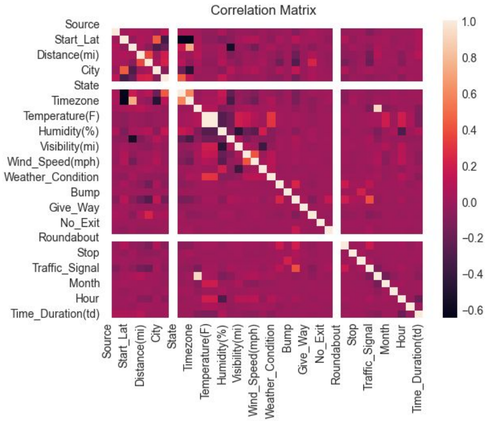
State-of-the-Art

Dr. Ronghui Zhou used a subset of the data to implement a prediction model. He removed any erroneous data and hand picked features for prediction. He used Random Forest, Decision Tree, Logistic Regression, and kNN with various k values and achieved a maximum accuracy of 97.4% with Random Forest with most accuracy scores being over 95%.

This is a very high prediction accuracy, and it was our goal to reach a similar accuracy score by testing various methods for each pipeline component.

Challenge

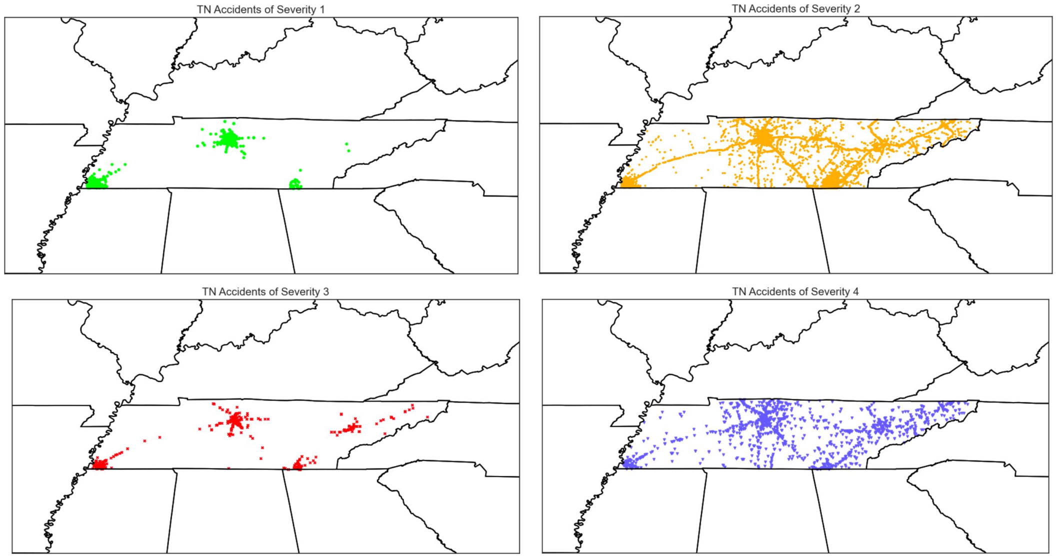
- There are 48 features and almost 8 million rows
  - This size makes it very difficult to process, so we had to think through how to effectively manage it
- Identifying important features for the prediction model was challenging
- Creating models with each dataset and classifier required a large amount of time across them



Tennessee Accident Severities

Because of the size of the complete dataset being over 3 GB, we decided to use a subset of the data for just the state of Tennessee

Prior to using this subset some of our tests ran for more than 3 hours for one classifier



TN Accidents of Severity 1

TN Accidents of Severity 2

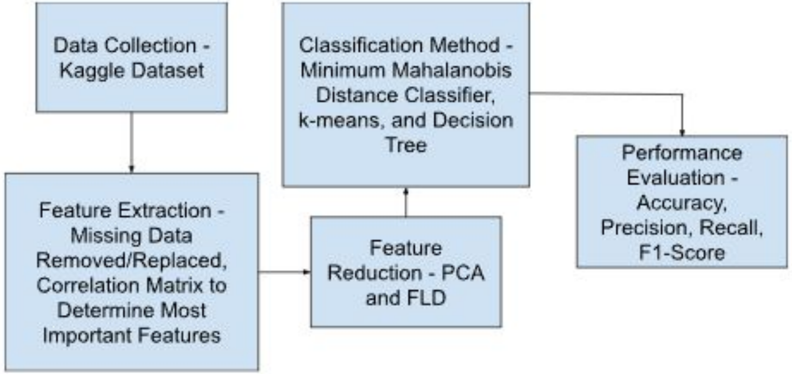
TN Accidents of Severity 3

TN Accidents of Severity 4

Algorithms

For this project, we used two methods for each step of the pipeline other than data collection

- Feature Extraction - Removed or replaced missing data and removed features based on a correlation matrix
- Feature Reduction - PCA and FLD
- Classification/Regression - Minimum Mahalanobis Distance Classifier, k-means, and Decision Tree
- Performance Evaluation - Accuracy, Precision, Recall, F1-Score


Data Collection - Kaggle Dataset

Feature Extraction - Missing Data Removed/Replaced, Correlation Matrix to Determine Most Important Features

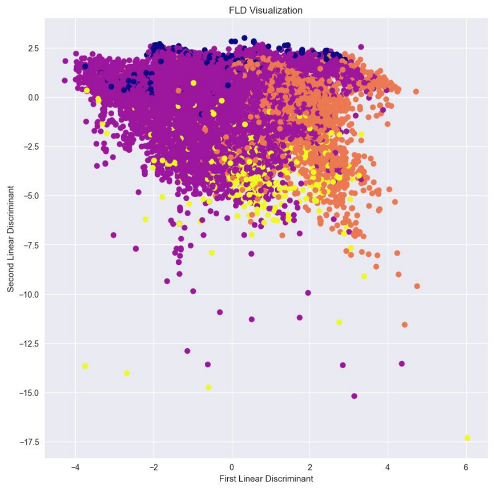
Feature Reduction - PCA and FLD

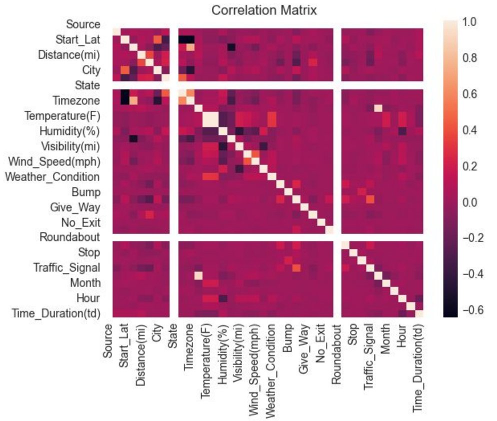
Classification Method - Minimum Mahalanobis Distance Classifier, k-means, and Decision Tree

Performance Evaluation - Accuracy, Precision, Recall, F1-Score

Design of Experiments

- We started with the raw data from Kaggle and began by creating new date/time columns and removed unnecessary columns such as ID. We then label encoded any string object fields to be integers for future steps
- We used a correlation matrix based on the 'Severity' value to determine the most important features and dropped the unimportant features from the data based on correlation or data distribution. This reduced the data to 36 features from 48 originally
- We split the data into 80% training, 10% testing, and 10% validation
- We applied PCA and FLD dimensionality reduction methods and reduced the dataset to 30 and 3 features, respectively, down from 6





Design of Experiments

- We used each classification method with the reduced datasets as well as the non-reduced dataset for comparison
  - The classification methods chosen were:
    - Minimum Mahalanobis Classifier (Bayesian, supervised, parametric)
    - K-means (Non-Bayesian, unsupervised, non-parametric)
    - Decision Tree (Regression)
- We applied Majority-Voting fusion to three similar accuracy models
- We also applied AdaBoost fusion with decision tree as the base estimator
- We reported the classification report for which includes: Accuracy, Precision, and Recall

Performance Evaluation

Classifier	Reduction Method	Accuracy	Precision	Recall
Minimum Mahalanobis Distance Classifier	PCA	30.91%	61%	31%
	FLD	48.40%	83%	48%
	None	48.79%	83%	49%
K-Means	PCA	28.57%	65%	29%
	FLD	34.69%	81%	35%
	None	21.38%	71%	21%
Decision Tree	PCA	54.66%	64%	55%
	FLD	76.25%	76%	76%
	None	88.88%	89%	89%
Decision Tree + AdaBoost	None	91.68%	91%	92%
Majority Voting	Maha. FLD/None and DT PCA	50.21%	83%	50%

Compared to State-of-the-Art

- Compared to the state-of-the-art, our efforts only achieved a maximum prediction accuracy of 91.68% whereas Dr. Zhou achieved 97.4%
- We used also used dimensionality reduction and fusion whereas Dr. Zhou did not
- Dr. Zhou also used a smaller subset of data from a single county in Pennsylvania, Montgomery County. This could explain the significantly higher accuracy on a much smaller subset of data
- He also used more regression-based methods than gaussian-based methods

Conclusions/Future Work

- The methods that we used did not necessarily work the best as our PCA accuracies were quite low
- We noticed that the dataset was not uniformly distributed between the classes of severity. There was significantly more entries of severity two accidents when compared to the other severities. (This hurt k-means accuracy most)
- Future work would involve further experimentation on which features are more correlated than others
    - This part was challenging due to the large number of features and relatively correlated features in the dataset
- Also, more investigation into regression techniques since those appear to have the best accuracy with this dataset