Interactive Visualization And Prediction Of U.S. Traffic Accidents

Zhengxi Qu, Guangwen Ren, Linna Su, Chang Xu, Qiaochu Zhu, Chao Chen

Team Project for CSE 6242 Georgia Tech, Spring 2022

Motivation

Why Important?

Traffic accident is one of the leading causes of death worldwide. Systems that can predict traffic accidents or accident-prone areas could potentially save lives.

Today's problem

- Today many map applications don't provide traffic accident predictions in advance.
- Academic studies that focus on predictions don't usually visualize their findings in an interactive way that public can easily understand.

Data

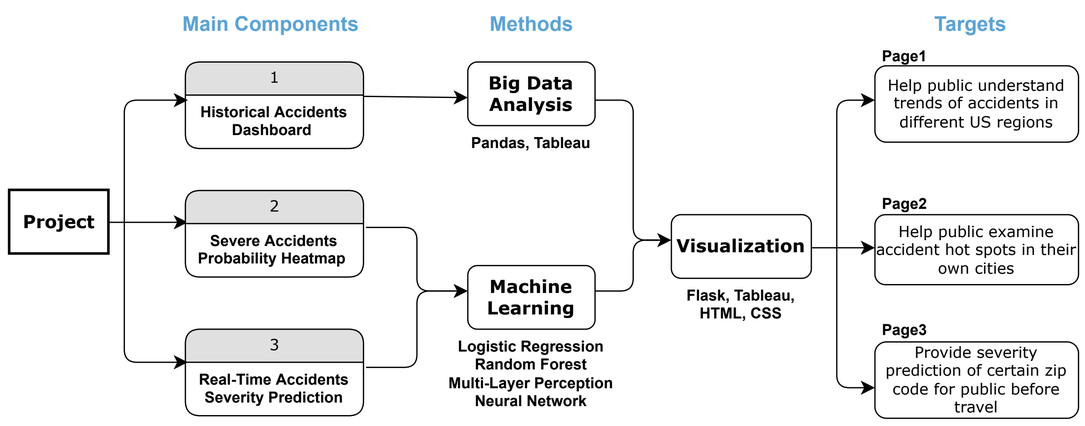
Accidents Data

The U.S. road traffic accident dataset downloaded from Kaggle has over **2.8 million** records from Jan 2016 to Dec 2021. Each record contains information including weather, location, time, severity of the accident, temperature, humidity, road conditions etc.

Real Time Weather Data

The real-time weather data is acquired from **OpenWeather API**. As users input a zipcode, recent weather data of that location will be returned and be used to predict accident severity.

Approach



Innovations

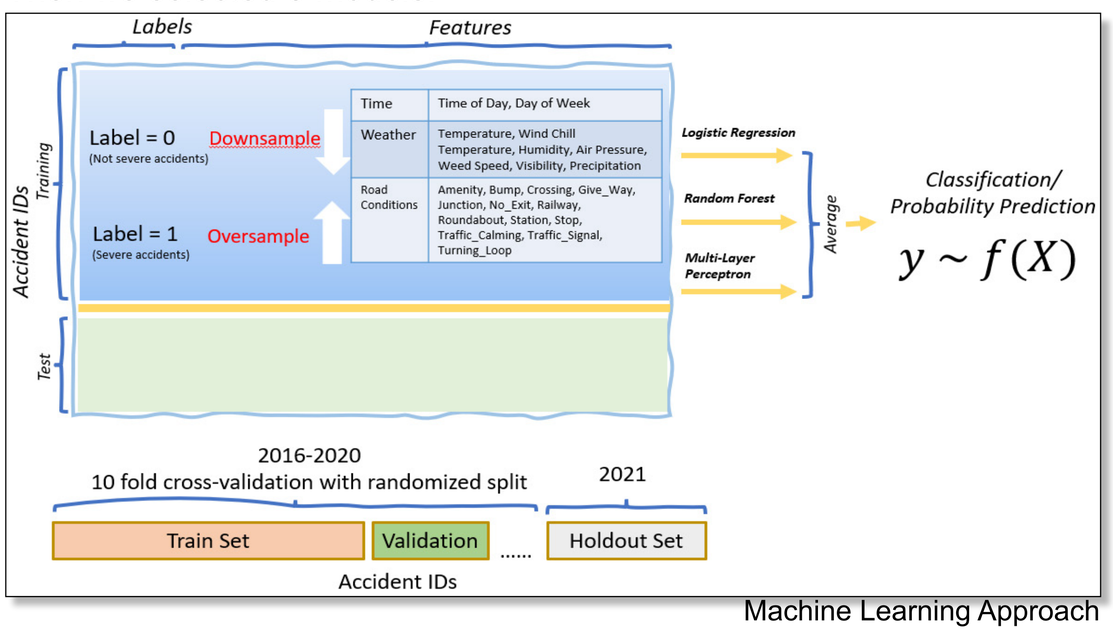
- Multiple machine learning models applied & tested
- Feature engineering and overfitting reduction techniques
- Historical accidents data visualized in interactive map
- Accidents prediction be fed with real-time weather data.
- Providing users the analytics and tools to avoid future traffic accidents and giving them assessment of the risk and early warnings in advance

Data Analysis and Visualization

For component 1 at country scale. We will process the 2.8M data records using Pandas and Tableau. Then generate a scalable interactive map of whole US showing accidents distribution/main causes.

Model Training & Validation

Since over 90% data points are non-severe accidents, we down sampled these data and over sampled the severe accidents data by bootstrapping. Then we selected 3 models.



Machine Learning Approach

- **Logistic Regression**: L2 regularization, parameter = 1/15
- **Random Forest**: n_estimators = 100, criterion for split 'gini', min_samples_split = 100, max_features = 0.8
- **Multi-Layer Perceptron Neural Network**: solver = stochastic gradient descent, hidden layer size = (100), activation function = 'relu', max iteration = 200

Experiments & Results

Model	Accuracy		ROC_AUC		F1	
	Train	Test	Train	Test	Train	Test
Logistic Regression	0.60	0.55	0.60	0.50	0.29	0.03
Random Forest	0.82	0.68	0.80	0.53	0.54	0.04
Multi-Layer Perceptron	0.87	0.98	0.51	0.50	0.03	0.01

Model Performance

Experiments

We experimented with various data transformation, feature engineering methods and model parameter setups.

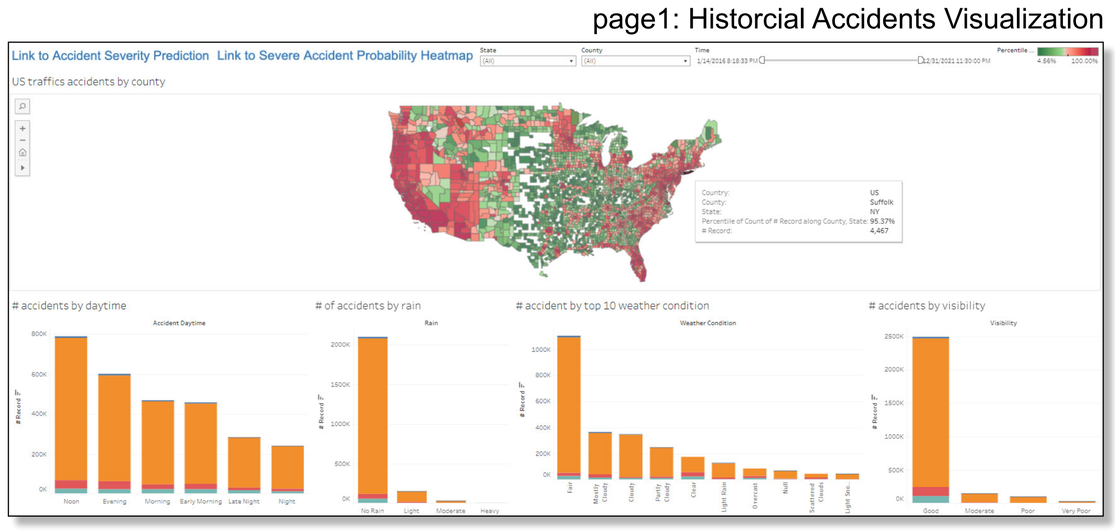
We use Accuracy (# of correct predictions/# of total predictions), AUC (Area Under Curve) of ROC Curve(Receiver Operating Characteristic Curve) to measure our model performance. An overall **prediction accuracy of 89% and AUC of 0.60** is achieved with consistent accuracy between train and test periods.

AUC and F1 score in test period are less satisfactory and prediction precision needs to be further improved.

Final Product & Evaluation

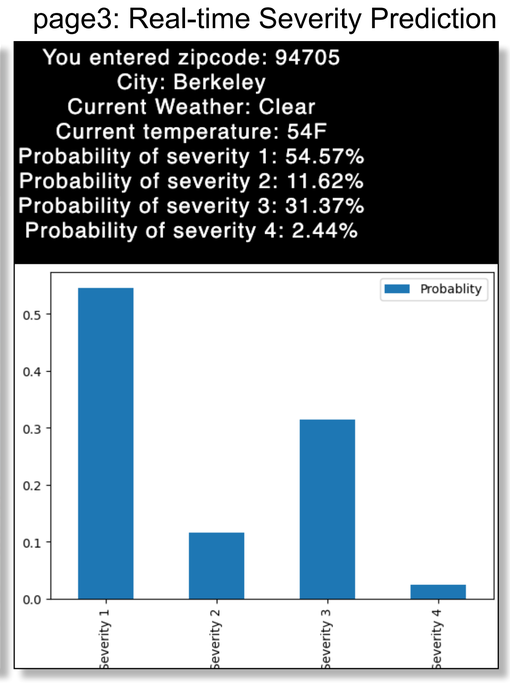
- 3-page web application.
- A user survey will also be conducted to assess success of the app.

page1: Historcial Accidents Visualization



page2: Historical Accidents Prediction

page3: Real-time Severity Prediction



Conclusion

- We combined 1 linear and 2 non-linear ML models and used various feature engineering, overfitting reduction techniques and achieved high predictoin accuracy.
- We designed a web application with interactive visualizations to help draw insights from historical data, and assess the risk of experiencing severe traffic accident in real time.
- Our product can be further improved by incorporating more relevant factors in real time and by being integrated with navigation applications to give travelers safety alert in real time.