Predicting California Car Accident Severity

December 5, 2021
Muyuan (Ernest) Liu | Duanduan Zhu
ml2699 | dz223

Abstract

Car accidents not only cause injuries and safety issues but also result in traffic distress. This project studies the causes of car crashes in California and predicts accident severity in terms of their impact on traffic. The project uses 1.5 million US car accident records and involves resampling, data engineering, missing value imputation, etc. The team then fit four supervised learning models using Logistic Regression, Random Forest, and Control Burn and select the best one based on accuracy and interpretability. Last, the team demonstrates model applications, propose three business uses, and discuss fairness and ethics of the model.

Table of Content

Table of Figures

1. Objectives and Values

The project has two objectives: understanding the causes of car crashes and predicting car accident severity. More specifically, we want to predict whether a given location in California will have a severe car crash based on time, weather, location, and traffic inputs.

Here, severity is defined in terms of accidents' impact on traffic and our team recognizes it to be of great value and interest. First, this is a new research area since there are many existing models that study the accident impact on personal injuries. Second, predicting accident impact on traffic can add business values to companies such as traffic applications, auto insurance, etc. Third, the public sector can prepare for or prevent severe accidents, so they can offer more efficient public services (e.g., city planning, EMT, etc.) and reduce economic loss due to bad traffic.

2. Preliminary Analysis

2.1 Data Introduction

Our dataset [1] contains 1.5 million car accident records in the US from 2016 to 2020. The dataset also consists of 47 features, the first one of which is the Y variable that we are predicting, "severity". As mentioned above, severity here is based on accidents' impact on traffic and takes values from 1 to 4. We will investigate this variable more closely later. The independent features in the dataset mainly come from 4 categories. **Time** features include the start and end timestamps of the accidents. **Weather** features include categorical descriptions of the weather, such as "snowy", "hailing", "southwest wind directions", etc. Weather features also record numerical values for the pressure, temperature, precipitation, etc. **Location** features contain the latitudes, longitudes, and street addresses of the accident site. Lastly, **point-of-interest** features describe traffic conditions of the site, such as if it is near a roundabout or a stop sign.

Too much: Drop Feature

Impute

Remove missing rows

All Feature Types and Missing Value Info		
	Missing Percentage	Data Type
Number	69.0%	float64
Precipitation(in)	33.7%	float64
Wind_Chill(F)	29.6%	float64
Wind_Speed(mph)	8.5%	float64
Humidity(%)	3.0%	float64
Visibility(mi)	2.9%	float64
Weather_Condition	2.9%	object
Temperature(F)	2.8%	float64
Wind_Direction	2.8%	object
Pressure(in)	2.4%	float64
Weather_Timestamp	2.0%	object
Airport_Code	0.3%	object
Timezone	0.2%	object
Zipcode	0.1%	object
Sunrise_Sunset	0.0%	object
Civil_Twilight	0.0%	object
Nautical_Twilight	0.0%	object
Astronomical_Twilight	0.0%	object
City	0.0%	object

Figure 1 Feature Types and Missing Value Information

2.2 Pre-Processing

Table 1 shows that the data set also contains missing data and we will treat them in 3 ways. The first feature, for instance, is missing roughly 70% of its values and should be **dropped entirely**. The last 8 features in Table 1 only miss less than 0.3% values and we simply **removed missing rows**. For the blue-shaded features that are missing a moderate amount of data, we will attempt to **impute** them later when we engineer them. We will then remove features with no real meanings, such as ID, and clean the data types into **only numeric and Boolean** values.

3. Define Focuses of the Problem

Despite having 5-year national car accident data, our team focused on predicting accident severity in California using data in California between July 2018 and July 2020 for the following reasons.

[1] (Moosavi, 2021)

3.1 Focus on California

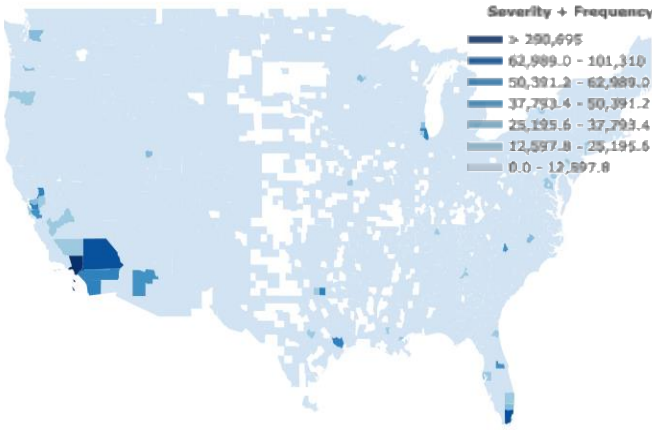

Figure 2 Map of US Accident Severity and Frequency (2016-2020)

In Figure 2, we plotted accident severity and frequencies in 1,671 counties. The dark blue shades indicate that severe accidents occur the most frequently in California, especially Los Angeles. And in fact, the distribution is **significantly skewed**, so most counties have very few severe accidents. There are **three reasons** why we chose to focus on California. First, other places have insufficient data, so choosing California can prevent overfitting. Second, California has different weather conditions and traffic policies than elsewhere (e.g. it does not snow), so there are intrinsic feature differences. Third, since there are high accidents frequencies, the demand to reduce accidents is also high.

3.2 Focus on July 2018 - July 2020

In Figure 3, we plotted the daily number of accidents in the past five years. We observed an unexpected decrease in accidents during July 2020, followed by a sharp increase and great volatility, which may be due to the COVID-19 pandemic. Data during this time are too volatile and have an abnormally higher mean. Moreover, data before 2018 would be somewhat outdated. To retain stable, sufficient, and recent data, we decided to keep data only from 2018-07-01 to 2020-07-01.

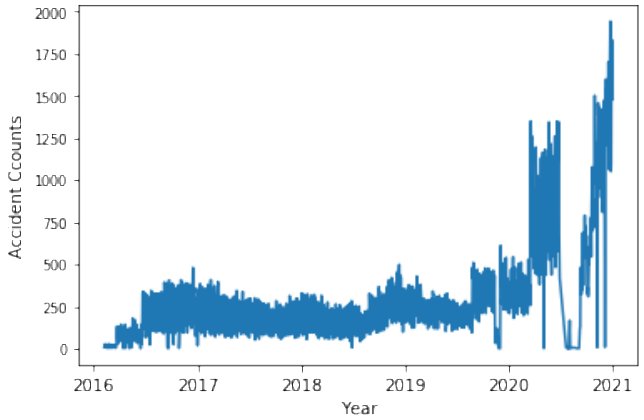

Figure 3 Daily Count of Accidents (2016-2021)

4. Data Engineering

4.1 Define Outcome Classes

Severity, the dependent variable, takes four integer values from 1 to 4. The challenge is that the original dataset gave no definition of the four severity levels, so all we know is that severity 1 is the lowest and that 4 is the highest. Therefore, we decided to define the severity ourselves by investigating the distance of the impacted traffic and the time duration of the impact.

Severity	Distance (mi)	Duration (min)
1	0.01	21.75
2	0.29	160.61
3	0.55	194.08
4	0.96	518.39

Figure 4 Definition of Severity Levels

From Figure 4, we observed that Severity 1 has very minimal impact on traffic. Since it only has less than 1% of total data, we dropped this class. Severity 2 and 3 have similar impacts to each other than to other classes and thus are combined into class "**Less Severe**". Severity 4 has the highest impact, 2 to 3 times larger than "Less Severe", so we renamed it as class "**More Severe**".
Our problem now becomes a **binary classification** problem.

4.2 Resampling

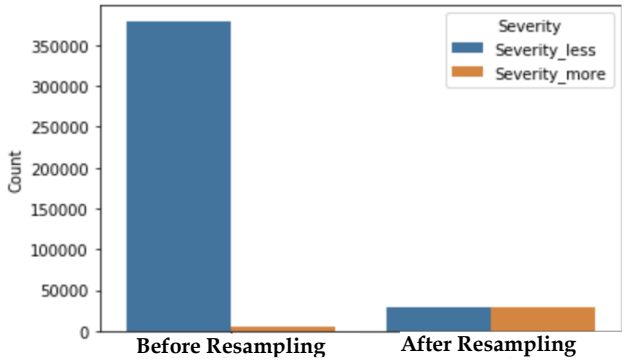

Figure 5 Sizes of the Two Outcome Classes after Resampling

After engineering the feature "severity", we find that the two classes are imbalanced, where 99% are "less severe", as shown on the left in Figure 5. We then performed resampling using the SMOTE method. We over-sampled the minority class to about 5 times its current value, and under-sampled the majority so that the two classes are balanced.

4.3 Time Features

We first look at the time features. Our problem is not a time series problem because we want to study the causes of severe accidents and time is only one of them. Currently, we have only one time feature in our original dataset, which is "Timestamp". Therefore, we broke down these timestamps into days, hours and weeks. After that, we plotted the accident severity frequencies by those metrics.

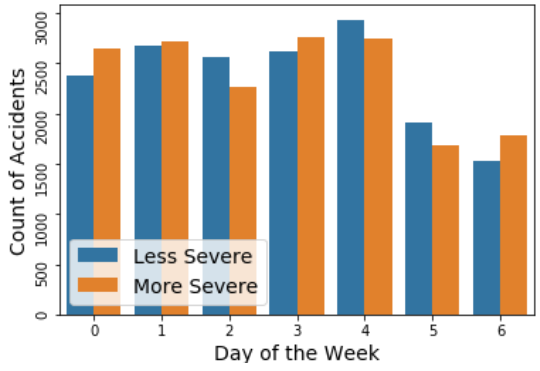

Figure 6 Counts of Accidents by Day of the Week

From Figure 6, we noticed that more accidents take place during weekdays, so we created a binary feature called "week_or_not."

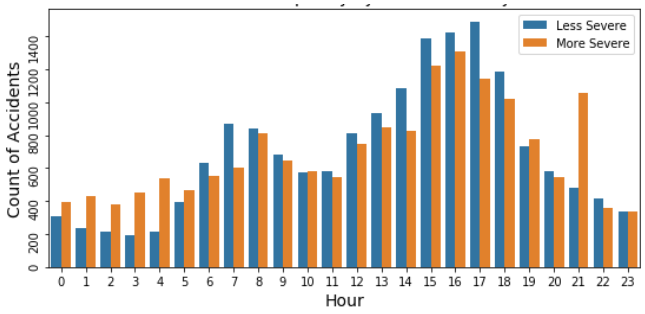

Figure 7 Count of Accidents by Hour of the Day

Figure 7 shows that there are more accidents during daytime, but more severe ones happen at night. We then created a binary feature, "day_or_night."

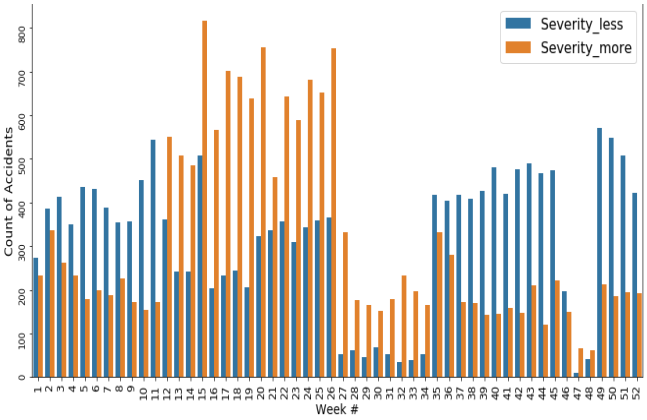

Figure 8 Count of Accidents by Week of the Year

From Figure 8, we observed some patterns due to seasons and holidays. For example, more severe accidents take place in spring and summer, and Thanksgiving has the least accidents in general.

It is worth noticing that feature "week" is indeed nominal because accidents are not impacted by the numeric values of weeks but the seasonal or holiday patterns. So we want to transform it into a categorical feature based on the patterns we observed. However, we face the challenge that there are some irregularities that we failed to explain and are hard to measure. For example, how do we measure the impact of holidays? Does Thanksgiving has the same impact as July 4th? Can we interpret all those patterns?

From Figure 8, we observed some patterns due to seasons and holidays. For example, more severe accidents take place in spring and summer, and Thanksgiving has the least accidents in general.

It is worth noticing that feature "week" is actually nominal because accidents are not impacted by the numeric values of weeks but the seasonal or holiday patterns. So we want to transform it into a categorical feature based on the patterns we observed. However, we face the challenge that there are some irregularities that we failed to explain and are hard to measure. For example, how do we measure the impact of holidays? Does Thanksgiving have the same impact as July 4th? Can we interpret all those patterns?

4.4 Impute Missing Values in Weather Features

As previously discussed in Pre-Processing, most missing values reside in weather features. We filled the missing data with values from the same city during the same week. We filled categorical features with the mode, the most frequent label, and filled continuous features with the median.

Moreover, we performed engineered texts in "weather conditions," a major weather indicator with 117 unique descriptions of the weather.

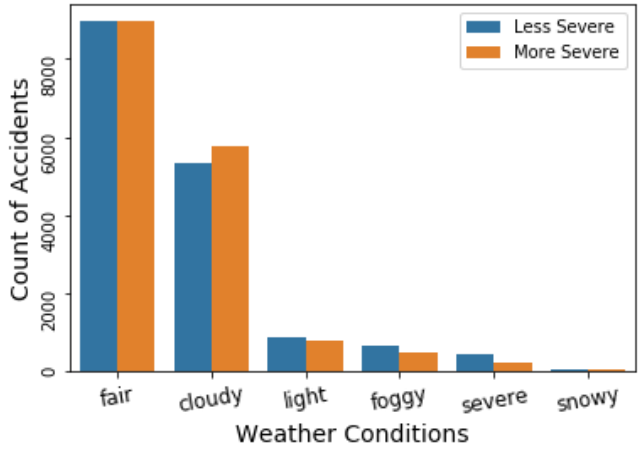

Figure 9 Count of Accidents by Weather Conditions

To reduce complexity, we applied text manipulation to sort the 117 labels into 6 groups, and extracted several insights, as shown in Figure 9. For example, severe accidents are more likely to happen on cloudy days and California generally has good weather. This result confirmed our decision to focus on California since the weather patterns here may not apply otherwise.

4.5 Eliminate Uninformative Features and One-Hot Encoding

There are two other kinds of features. Coordinates of the accidents can represent location very well and contain no missing values. Point-of-interest features, along with some other minor features, lack variation and do not yield useful insights. We will drop such features.

After applying one-hot encoding, we obtained a balanced dataset with 40,000 observations and 20 features, all of which take continuous or Boolean values.

5. Model Fitting and Selection

We implemented 4 models sequentially. We first tried Logistic Regression and Random Forest using all 20 features, in which we used grid search to find the best hyperparameters and used train/test split (8:2) and cross-validation to avoid overfitting. Then, we utilized Random Forest and Control Burn to select fewer features. Model results will be compared and discussed together.

5.1 Logistics Regression

We first used Logistic Regression, which is the most classic model for a **binary classification** problem like this. After standardizing and applying LASSO regularization, the model achieved **only an accuracy** of 0.60 in both outcome classes and in both train and test datasets. The unsatisfactory performance suggests non-linearity in the dataset and encourages us to switch to tree-based models.

5.2 Random Forest with 20 features

As discussed above, the 52-category nominal feature "week" and the non-linearity in the dataset can be better addressed by a tree-based model. Plus, Random Forest is a more robust ensemble learner and can help choose important features. The model achieved an adequate test accuracy of 0.97 (cross-validated and weighted) and obtained the following feature importance rankings.

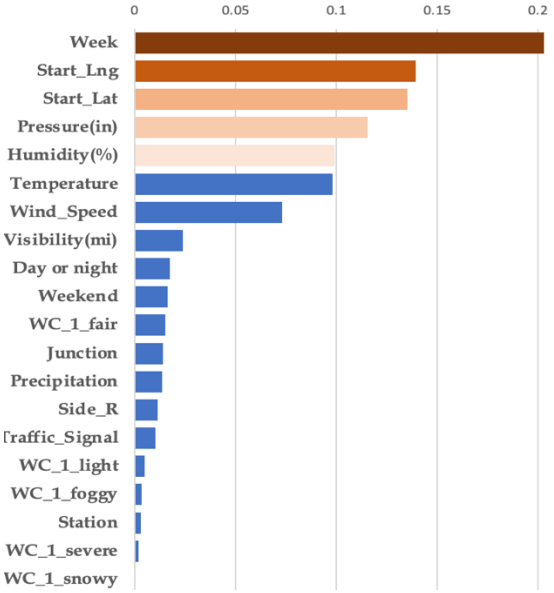

Figure 10 Importance of 20 Features in Random Forest

From Figure 10, we can observe that the last few features have little importance and could be eliminated. Their low significance partially agrees with the weather conditions in California, since it barely snows or has severely wet weather there. On the other hand, humidity and pressure are usually affected by rainy weather, which can affect traffic in California and thus are important features. Other weather features are similar to these two. We concluded that the top 5 features account for 70% of total importance and should be studied further.

5.3 Random Forest with 5 Features

We constructed a simpler random forest with only the top five features from before. They are week, coordinates, pressure, and humidity that describe the time, location, and weather conditions. The model obtained a similar good accuracy of 0.96, compared to 0.97 in the all-feature random forest model. We concluded that these 5 features could explain the causes of severe car accidents and make a mostly accurate prediction.

5.4 Control Burn

As the last step, we utilized Control Burn, which is similar to Random Forest as it is also a tree-based ensemble learner. But more importantly, it can help select important features since "it uses a weighted LASSO-based feature selection method to prune unnecessary features from tree ensembles, just as low-intensity fire reduces overgrown vegetation[2]."

Control Burn Results and Parameters		
Regularization	Accuracy	# of Feature Selected
0.0045	0.6	1
Feature Selected: 'Week'		
0.0042	0.69	2
Feature Selected: 'Week' , 'Day/Night'		
0.004-0.002	0.96	10
Feature Selected: 'Week' , 'Day/Night', 'Lng', 'Lat', etc.		
0.002	0.96	20
Feature Selected: All		

Figure 11 Control Burn Results and Parameters

Table 11 shows accuracy and feature selection results according to regularization strength. All models selected "week" as the most (and sometimes sole) important feature, which confirmed our results from Random Forest.

However, Control Burn is unable to select fewer features while retaining good accuracy. In order to achieve a similar score of 0.96 as Random Forest, the model selected at least half of all features. This result is less desirable than Random Forest.

6. Model Results

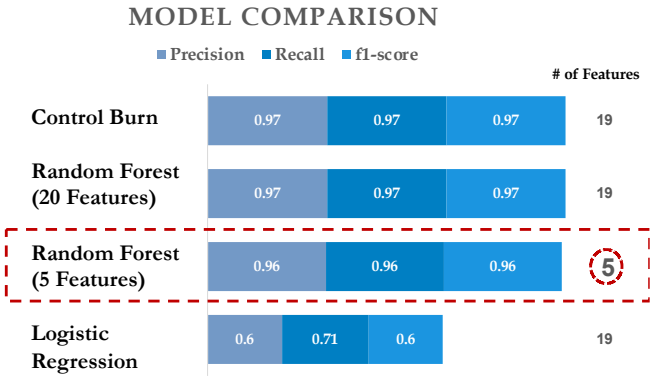

Figure 12 Summary of Model Performances

In conclusion, Random Forest with 5 Features is the best model because of its high accuracy and interpretability. It achieved a 0.96 accuracy score using only 5 features.

We are confident of this result because we have taken several steps to avoid over- and under-fitting. We engineered each feature, eliminated those with little variance and little information, and included features of multiple meanings and aspects. Second,

[2] (Brian Liu, 2021)

we have a concise set of only 20 features and a sufficiently large and balanced set of 40,000 observations. Lastly, we selected tree-based models and implemented train test split and cross-validation for the results.

7. Applications and Potential Business Uses

7.1 Model Application Example

We will now demonstrate how to apply the best-performing the 5-Feature Random Forest model to **predict accident severity in Los Angeles through a synthetic dataset**.

The model has 3 kinds of features: week, coordinates (latitudes and longitudes), and weather (pressure and humidity). First, we can fix the **week** number in the model to generate weekly forecasts. In the example dataset, we chose the 27th week in a year. Second, we can enter any **coordinate** pairs in California into the model. Our demo dataset first used known coordinates from the existing dataset (Figure 13), and then generalizes the results through a uniform grid of coordinates in LA (Figure 14). Last, the model can take inputs from **weather** forecasts and in the demo, we used historical median values.

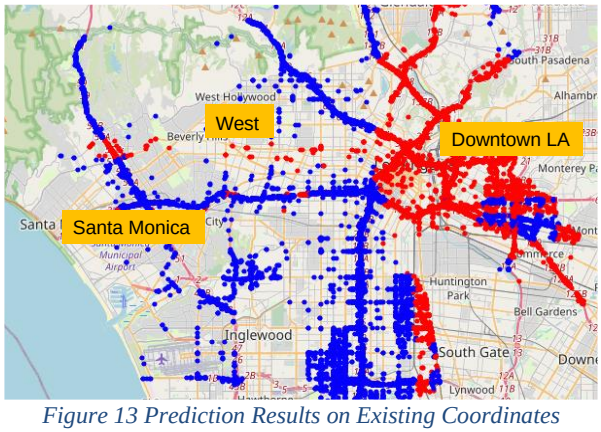

Figure 13 Prediction Results on Existing Coordinates

Figure 14 Prediction Results on Synthetic Coordinate Grid

In both figures, red denotes areas that are predicted to have more severe accidents more likely than blue-shaped areas. It appears that Downtown LA, West Hollywood, and Beverly Hills will have more severe accidents than other places such as Santa Monica. The coordinate grid (Figure 14) better partitioned the areas based on the prediction.

7.2 Three Potential Business Use Cases

Predicting accidents with a severe impact on traffic can add value to businesses. We now propose 3 use cases. **Traffic applications** can incorporate this model to predict places that are prone to more severe accidents and warn their users to avoid those areas. This use brings profitability for software companies and convenience to the **general public**.

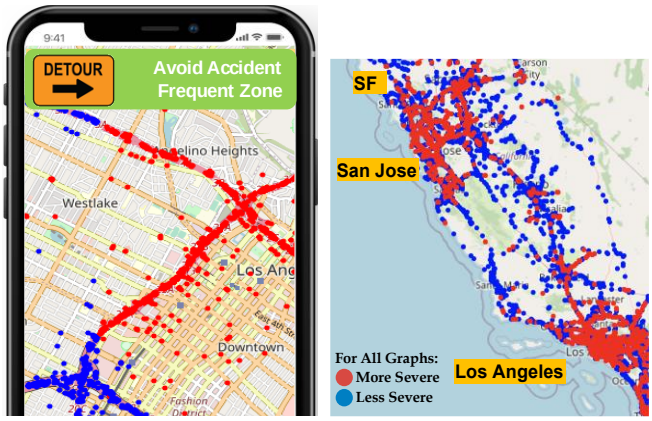

Figure 15 Potential Use Cases 1&2

Auto insurance companies can adjust policy premiums based on where and when severe accidents may happen. In Figure 15, our model predicted Los Angeles, San Jose, and San Francisco to have more severe cases, and insurance there may have a little higher price. The model helps companies make decisions in a data-driven manner.

Figure 16 Potential Use Cases 3 (DTLA Map)

Public sectors can provide more efficient services and reduce economic loss due to bad traffic. They can better **prepare for the accidents** using model prediction. Figure 16 is a map of LA, where Downtown seems to have more severe accidents (red dots) than surrounding areas. So EMTs can prepare backup routes should the traffic there be severely impacted due to accidents. In the upper left corner, we can see that severe accidents might occur next to the freeway exit, so city planning can put up signs and warnings for drivers, in order to **prevent accidents** from happening.

8. Fairness and Ethics

Our model is fair and not a Weapon of Math Destruction.

8.1 Fairness

Since our dataset contains no driver information, there would be no discrimination based on age, gender, race, or health conditions. Moreover, all of our input features, including weather, location, or time, are objective.

8.2 Weapon of Math Destruction

Our model is not a WMD. First, our outcomes are not hard to measure because the accident data could be collected constantly, and new data could be used to justify and modify our current model. Second, our model does not generate negative consequences and does not harm anyone. Since no predictions are 100%, there could be some risks, where inaccurate predictions may cause users to detour, insurance companies to lose money, and public services to falsely allocate resources. However, even weather forecast inputs could be wrong, the loss is reasonable and controllable.

Third, our model does not create feedback loops. Since our inputs are objective, the predicted results cannot affect the next round inputs, so there would be no feedback loops.

9. Bibliography

Brian Liu, M. X. (2021, July 1). *ControlBurn: Feature Selection by Sparse Forests.* Retrieved from arXiv.org: https://arxiv.org/abs/2107.00219

Moosavi, S. (2021, January 1). *US-accidents: A countrywide traffic accident dataset.* Retrieved from https://smoosavi.org/datasets/us_accidents