Austin Traffic Incidents in Recent Years (2017-2022)

By Darren Au, Jared Pomerantz, Arturo Hernandez, and Dao Ton-Nu

Overview

Dataset: Records Traffic Incidents occurring in Austin. The sample taken includes the years 2017-2022. We cleaned the sample data by removing records where longitude or latitude were inputted as 0 (not in Austin, Texas).

Research Question: We are looking at the common features of traffic issues in Austin and the classification of incidents. We want to find the best classifier to predict, based on given features, whether a traffic incident falls during weekdays or weekends. Secondarily, we also wanted to find the best classifier to predict the type of incident.

Overview (cont.)

Initial Expectations: We expect that there won't be a very strong accuracy or overall performance with our classifiers. However, we expect there may be some pattern with when incidents happen that we can see, either through the classifiers or through general distribution observations.

Method of Evaluating Project: We evaluate the performance of our classifiers using accuracy and a classification report including precision, recall, and f1-scores. For our best performing classifiers, we also took into account the number of true and false positives.

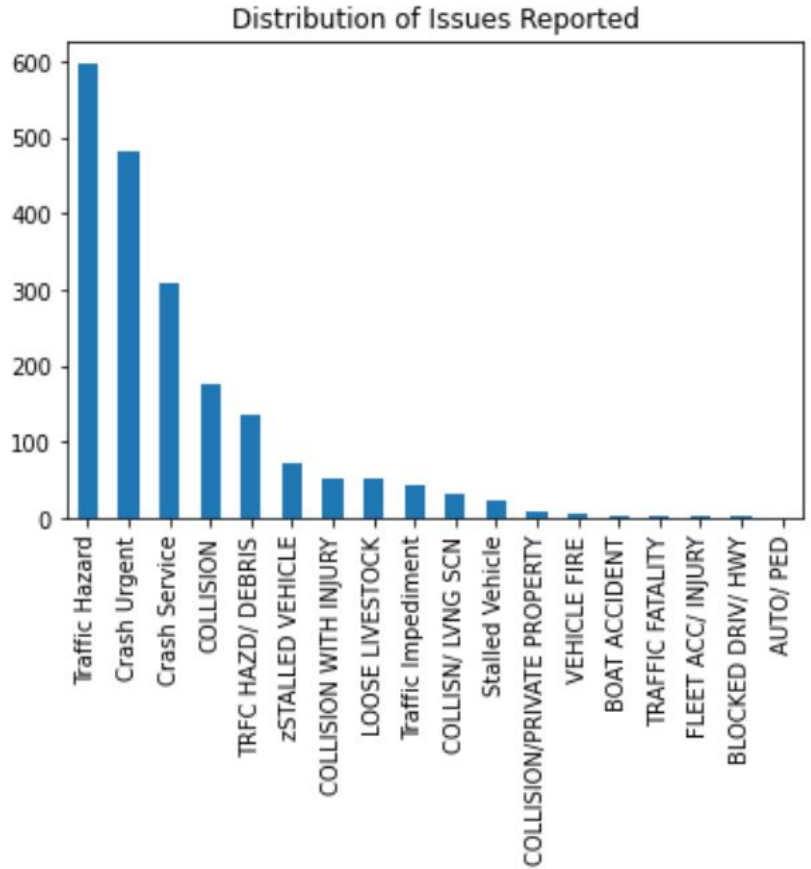
Data Visualization and Distributions

Distribution of Issues Reported

(by Darren)

The issue category reported the most in our sample was for Traffic Hazards. Most of the bigger categories are related to hazards, crash, or collision.

For the issue-type classification algorithms, we would eventually focus on two main groups: Hazards and Collisions.

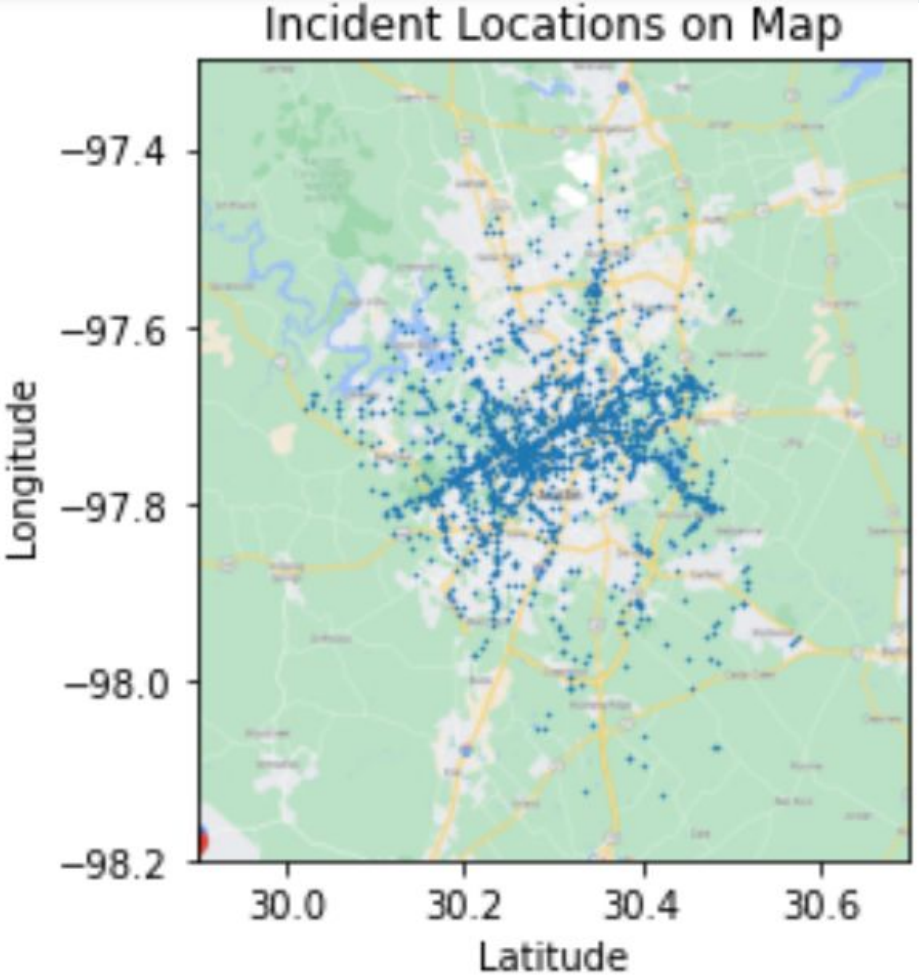


Location of Incidents

(by Jared)

As can be expected, most traffic incidents are on or near streets and in the populated area.

Notably, there is more concentration of incidents near streets such as highways and freeways, where the speed limits are higher than in residential or school roads.

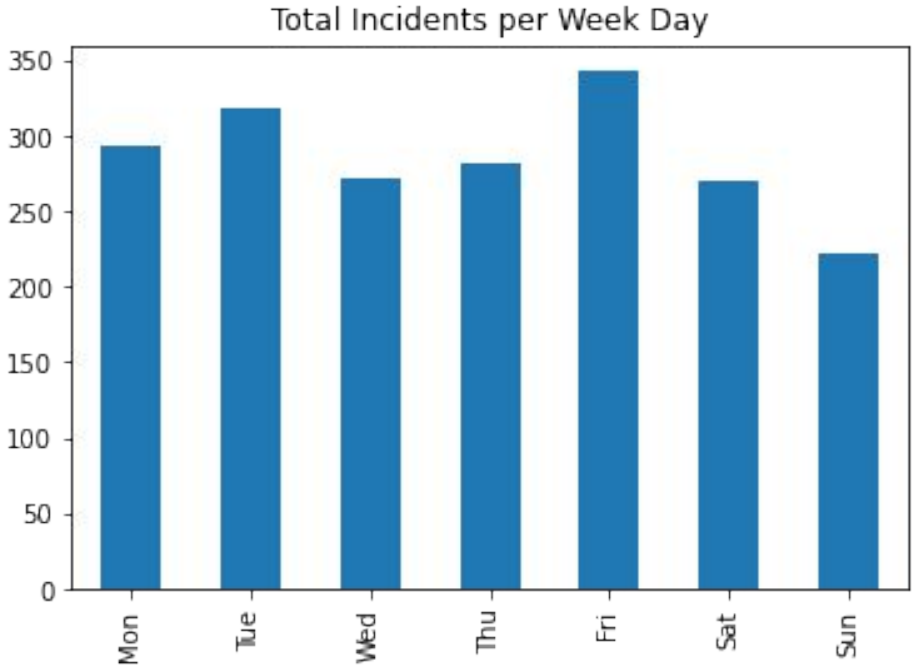


Day of Week Distribution

(by Arturo)

The day of the week with the most incidents was Friday, with 343 incidents. All the weekdays had more accidents than Saturday and Sunday.

For the classification algorithms, we will predict whether an incident occurs on a weekday or a weekend.

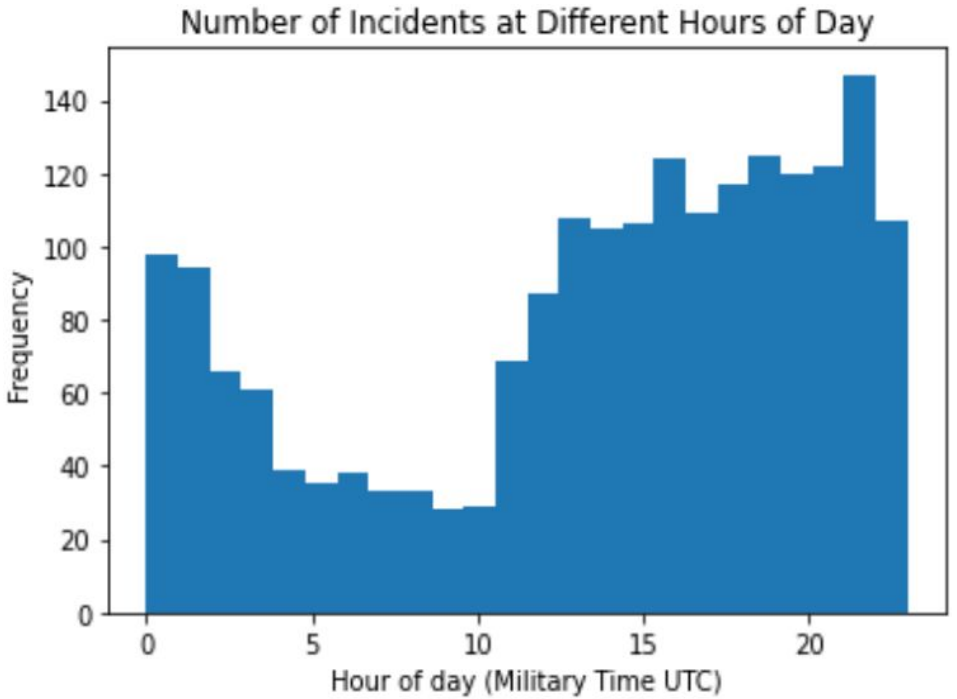


Incident Hour of Day Distribution

(by Dao)

The time values were obtained by simplifying the dataset's times into hour values in military time (UTC, converted to CST for observations).

Looking at the distribution, the time with the least incidents is at about 3-4 AM CST. The time with the highest incidents is 4 PM CST.



Prediction Algorithms

Predictions/Results

To make classification predictions, we wrote a script that would try each of the six classification methods (logistic regression, random forest, decision tree, KNN, SVM, Adaboost), and produce the accuracy and classification report for each of them.

Weekday vs Weekend Predicting

From the output file, decision tree performs significantly worse while all the other classification algorithms perform similarly. Adaboost marginally performs the best out of all the algorithms with an accuracy of 0.7825.

However, most of the algorithms have a very low recall score for weekends, meaning when the true value is a weekend, the algorithms falsely predict it as a weekday a very high percentage of the time.

logistic regression
0.78

	precision	recall	f1-score	support
0	0.00	0.00	0.00	88
1	0.78	1.00	0.88	312
accuracy			0.78	400
macro avg	0.39	0.50	0.44	400
weighted avg	0.61	0.78	0.68	400

random forest
0.755

	precision	recall	f1-score	support
0	0.36	0.15	0.21	88
1	0.79	0.93	0.86	312
accuracy			0.76	400
macro avg	0.58	0.54	0.53	400
weighted avg	0.70	0.76	0.71	400

decision tree
0.655

	precision	recall	f1-score	support
0	0.26	0.31	0.28	88
1	0.79	0.75	0.77	312
accuracy			0.66	400
macro avg	0.53	0.53	0.53	400
weighted avg	0.68	0.66	0.66	400

KNN
0.78

	precision	recall	f1-score	support
0	0.50	0.01	0.02	88
1	0.78	1.00	0.88	312
accuracy			0.78	400
macro avg	0.64	0.50	0.45	400
weighted avg	0.72	0.78	0.69	400

svm
0.78

	precision	recall	f1-score	support
0	0.00	0.00	0.00	88
1	0.78	1.00	0.88	312
accuracy			0.78	400
macro avg	0.39	0.50	0.44	400
weighted avg	0.61	0.78	0.68	400

adaboost
0.7825

	precision	recall	f1-score	support
0	0.57	0.05	0.08	88
1	0.79	0.99	0.88	312
accuracy			0.78	400
macro avg	0.68	0.52	0.48	400
weighted avg	0.74	0.78	0.70	400

Issue Type Predicting

Our initial attempt was to see if the prediction algorithm could determine among all the different issue types. However, we were receiving a very low accuracy due to there too many issue types and thus, decided to take the dataset we had and only select the the issue types that would fall into the categories of hazards and collisions. The size of this dataset is 1098.

Hazards
'OBSTRUCT HWY', 'BLOCKED DRIV/ HWY', 'Traffic Impediment', 'LOOSE LIVESTOCK', 'TRFC HAZD/ DEBRIS'
‘Traffic Hazards’

Collisions
'COLLISION WITH INJURY', 'COLLISN/ LVNG SCN', 'COLLISION/PRIVATE PROPERTY',
‘COLLISION’

logistic regression
0.7410714285714286

	precision	recall	f1-score	support
COLLISION	0.00	0.00	0.00	58
Traffic Hazard	0.74	1.00	0.85	166
accuracy			0.74	224
macro avg	0.37	0.50	0.43	224
weighted avg	0.55	0.74	0.63	224

random forest
0.8035714285714286

	precision	recall	f1-score	support
COLLISION	0.62	0.62	0.62	58
Traffic Hazard	0.87	0.87	0.87	166
accuracy			0.80	224
macro avg	0.74	0.74	0.74	224
weighted avg	0.80	0.80	0.80	224

decision tree
0.75

	precision	recall	f1-score	support
COLLISION	0.52	0.57	0.54	58
Traffic Hazard	0.84	0.81	0.83	166
accuracy			0.75	224
macro avg	0.68	0.69	0.68	224
weighted avg	0.76	0.75	0.75	224

KNN
0.7410714285714286

	precision	recall	f1-score	support
COLLISION	0.00	0.00	0.00	58
Traffic Hazard	0.74	1.00	0.85	166
accuracy			0.74	224
macro avg	0.37	0.50	0.43	224
weighted avg	0.55	0.74	0.63	224

svm
0.7410714285714286

	precision	recall	f1-score	support
COLLISION	0.00	0.00	0.00	58
Traffic Hazard	0.74	1.00	0.85	166
accuracy			0.74	224
macro avg	0.37	0.50	0.43	224
weighted avg	0.55	0.74	0.63	224

adaboost
0.7366071428571429

	precision	recall	f1-score	support
COLLISION	0.49	0.41	0.45	58
Traffic Hazard	0.81	0.85	0.83	166
accuracy			0.74	224
macro avg	0.65	0.63	0.64	224
weighted avg	0.72	0.74	0.73	224

From the output file, random forest performs slightly better with an accuracy of 0.80 while all the other classification algorithms perform similarly. Collisions performed worse than traffic hazard with an f1 score of 0.62 vs 87

Conclusion

Overall, although we were able to visibly observe some patterns, our algorithms were not successful, as we had anticipated.

The algorithms do a decent job of predicting, but could definitely use improvement. It would probably benefit from a larger dataset as well as more parameters. What surprised us was the lack of differences in the prediction algorithms, as the dataset was presumably non linear, and therefore, logistic regression and SVM should struggle compared to the other algorithms. However, our assumption was once again that this was attributed to the lack of training given and these differences would be more stark as the models are fed more data.