Weather Impact On Accident Traffic Severity

Springboard Capstone Project 1
Greg Gibson June 2020

Contents

1. Problem Statement

Can the severity of traffic, post accident, be predicted using weather metrics?

Upon occurrence of a traffic accident, can local weather conditions, time of day and other related factors combine to predict the severity of the traffic delay? Travel route applications, whether on smartphones, provided by portable GPS units, or contained within vehicle dashboards, can utilize position tracking of various motorists to indicate how much additional time it may take to travel past the problem, and those who are able will find alternate routes. But, if an algorithm could predict how long the traffic may persist, the application could suggest travelers simply wait out the extra minutes or consider how long to delay before starting their trip. This tool could also be adapted to all traffic root-cause types such as flat tire change, police presence, construction, sporting event, etc. An expansion of routing application services would attract additional users.

2. Clients

Who would be interested in providing such information? Popular smartphone routing applications are available by Apple and Google, as well as downloadable apps such as those developed by Waze and Mapquest. Portable GPS units are led by Garmin, and in-dash navigation systems are produced by car-audio companies such as Boss, Kenwood, and Alpine.

In addition, weather stations providing updates on the radio or through their own smartphone applications can also assist motorists by predicting accident traffic severity with their own weather data.

3. Data

The data used, as explained on the following pages, has information on weather conditions at the time of numerous vehicle accidents. There are many other factors potentially involved in how severe traffic can be after an accident - the number of vehicles involved, the number of lanes and speed limit, one or two directional, extent of injuries, rural to urban location, etc. These other potential factors are not included in this analysis, but if weather has a proven correlation, it should be one of the factors in a more comprehensive algorithm.

Dataset:

Source - https://smoosavi.org/datasets/us_accidents

File - US_Accidents_Dec19.tar.gz

Acknowledgement - Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, arXiv preprint arXiv:1906.05409 (2019).

Website’s Description - This is a countrywide traffic accident dataset, which covers 49 states of the United States. The data is continuously being collected from February 2016, using several data providers, including two APIs which provide streaming traffic event data. These APIs broadcast traffic events captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about 3.0 million accident records in this dataset.

There are 48 features, including:

- A 1-4 ranking of Severity, from least to most significant impact to traffic delay
- Length of traffic in miles
- Start and end timestamps and day/night indicators
- Address information plus GPS coordinates
- Airport weather station code nearest accident
- Eight weather metrics such as temperature, precipitation and visibility, plus description
- Thirteen points of interest such as traffic signals, intersections and crosswalks

Wrangling:

A couple features pertaining to location were frequently blank, ~65% - 75%, and discarded due to the focus on the weather and potential difficulty to recreate, these were the end point GPS of the traffic and the street number.

Two features necessary for this analysis, Precipitation and Wind Chill, were similarly blank. Attempts were made to fill the information through APIs but abandoned due to record count limitations. Alternatively, since there is multi-year weather in the database, the assumption is similar weather each month per region, i.e. New England April showers or Arizona July dry heat. The data was sorted by month and zip code prefix, and filled these blanks with interpolation.

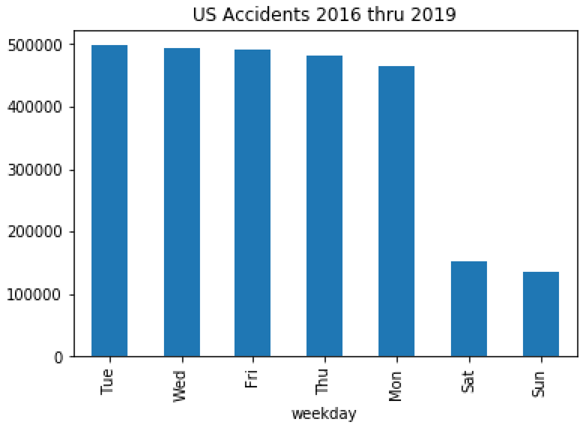
Another feature removed, TMC, Traffic Message Channel, has over 2,000 descriptions of accidents. It is 25% blank in the database, and the balance is 65% code 201 for “accident”.

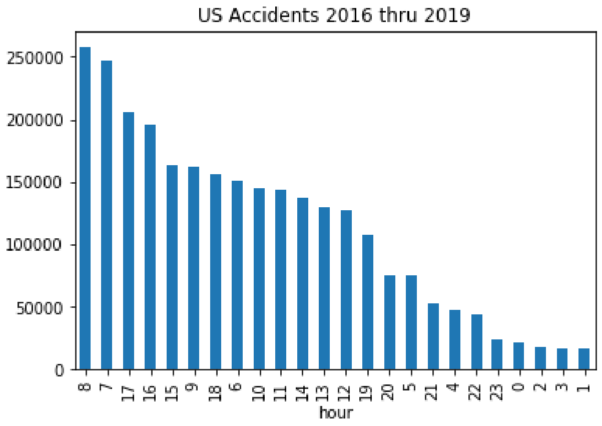
Outliers were excluded using a stats.zscore threshold of three standard deviations on type float columns.

Nine columns had about 2% or less blank, and due to the large number of records near 3M, these rows were dropped without research, retaining 91% of the original large record count.

4. Exploratory Data Analysis (EDA)

There are likely some common assumptions about traffic accidents, such as when they occur and whether increasing or decreasing. To check the dataset is reasonable at a high level, a few simple comparisons were graphed.

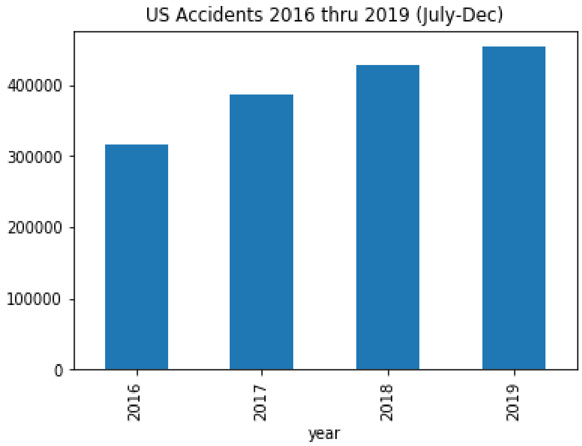


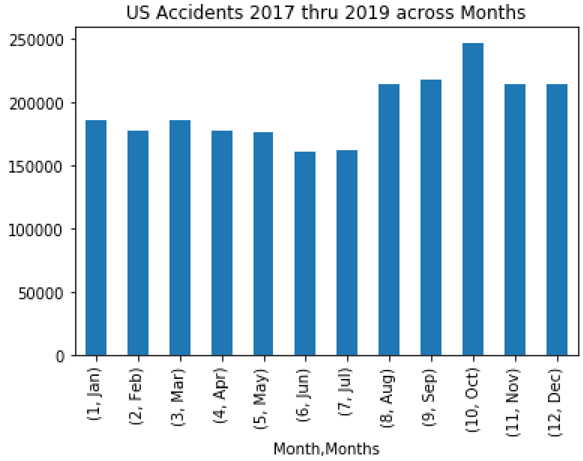


More accidents and similar counts reported on weekdays, rather than weekends, and more accidents reported during the morning and afternoon commuting hours. Logically, most accidents would happen when people are driving to work.

There is also a visible trend in yearly accidents. This may coincide possibly with more distracted drivers due to smartphone usage, or as unemployment rates improved during this time period, more people drove to work.

(Note: Data appears partial until June in 2016 and all years measured over the same months.)

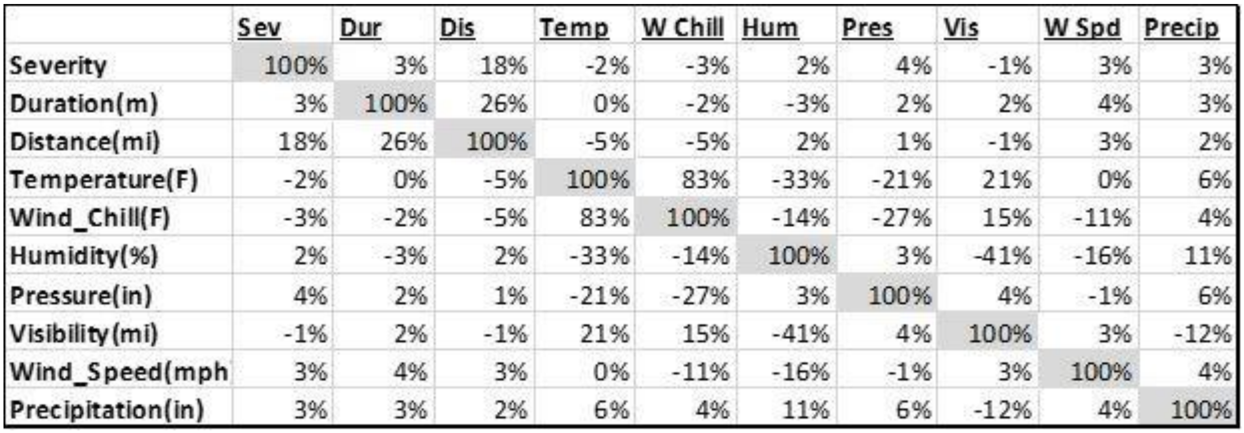




Out of curiosity, how do accidents look by month? June & July summer travel is low as vacations are spread out, no school schedule, and peak sunlight.

The last five months of the calendar year average 23% higher than the first five months. October, with Columbus Day weekend, Halloween, and a time change to darker commute hours, has the highest number of accidents.

Correlation Matrix:



	Sev	Dur	Dis	Temp	W Chill	Hum	Pres	Vis	W Spd	Precip
Severity	100%	3%	18%	-2%	-3%	2%	4%	-1%	3%	3%
Duration(m)	3%	100%	26%	0%	-2%	-3%	2%	2%	4%	3%
Distance(mi)	18%	26%	100%	-5%	-5%	2%	1%	-1%	3%	2%
Temperature(F)	-2%	0%	-5%	100%	83%	-33%	-21%	21%	0%	6%
Wind_Chill(F)	-3%	-2%	-5%	83%	100%	-14%	-27%	15%	-11%	4%
Humidity(%)	2%	-3%	2%	-33%	-14%	100%	3%	-41%	-16%	11%
Pressure(in)	4%	2%	1%	-21%	-27%	3%	100%	4%	-1%	6%
Visibility(mi)	-1%	2%	-1%	21%	15%	-41%	4%	100%	3%	-12%
Wind_Speed(mph	3%	4%	3%	0%	-11%	-16%	-1%	3%	100%	4%
Precipitation(in)	3%	3%	2%	6%	4%	11%	6%	-12%	4%	100%

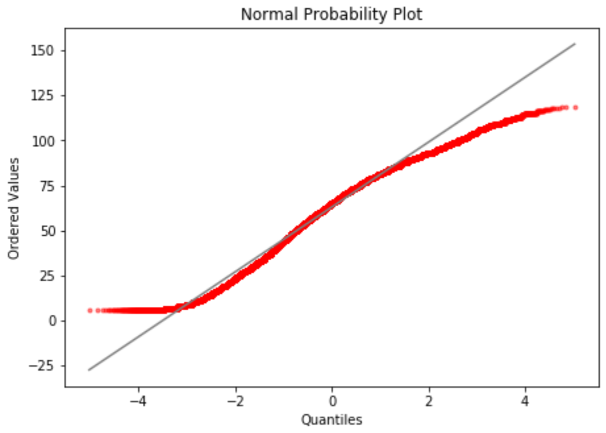
A correlation matrix shows no correlation between the traffic Severity score and the weather features with most rates near 0%. Only Distance is weakly correlated.

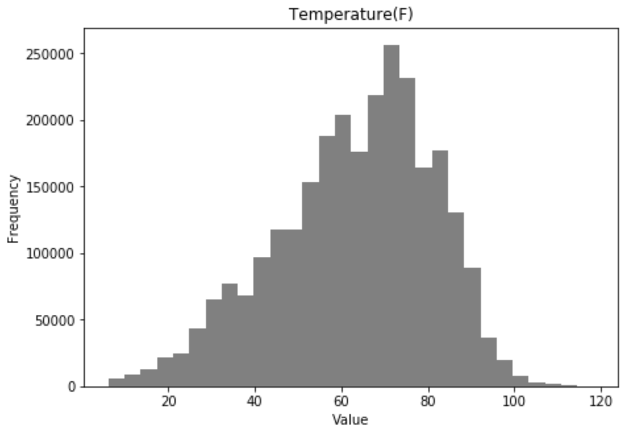
The only strong correlation is between Temperature and Wind Chill, which is logical due to the experience in the winter months. Of interest, a moderate, negative correlation between Visibility and Humidity, as increased humidity can create foggy or rainy driving conditions. Temperature and Humidity are also negatively correlated. According to online sources, as air temperature increases, air can hold more water vapor, and its relative humidity decreases.

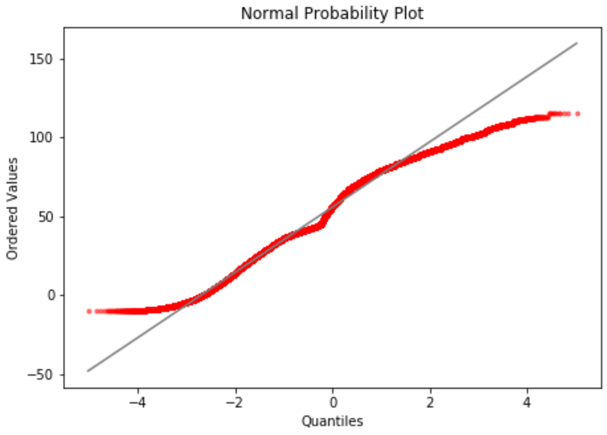
5. Inferential Statistical Analysis

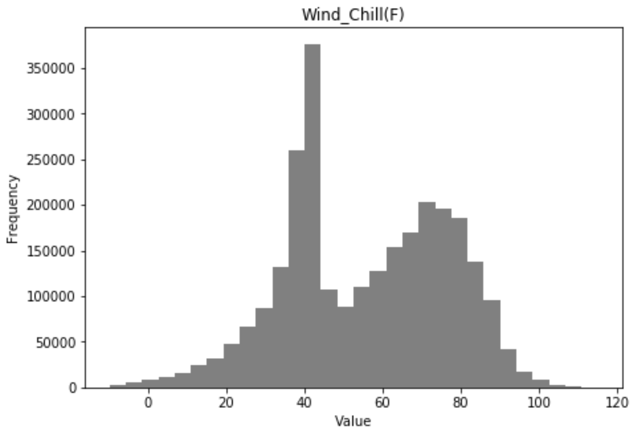
The first issue to note is that none of our feature distributions are normal. Below are examples of QQ plots with various departures from normality.

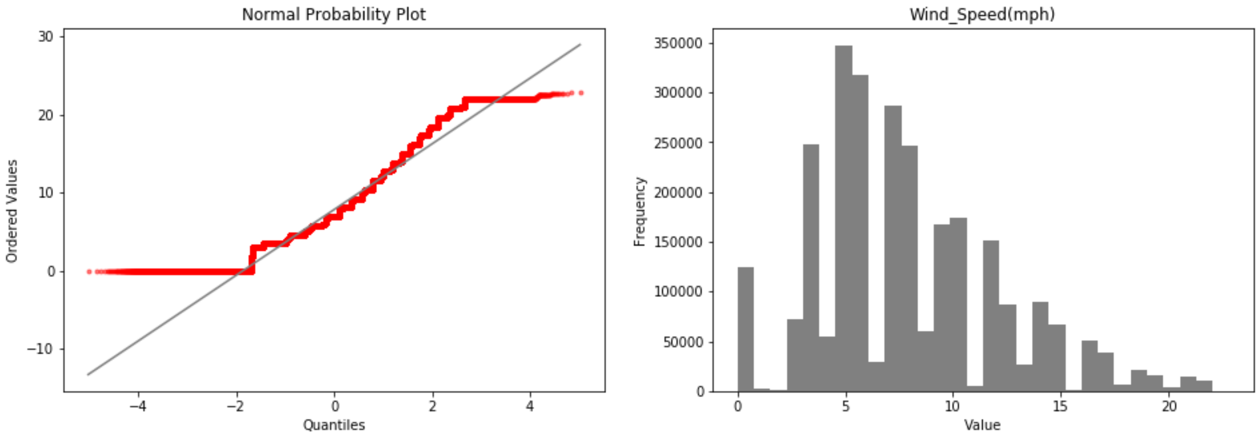
Distributions:



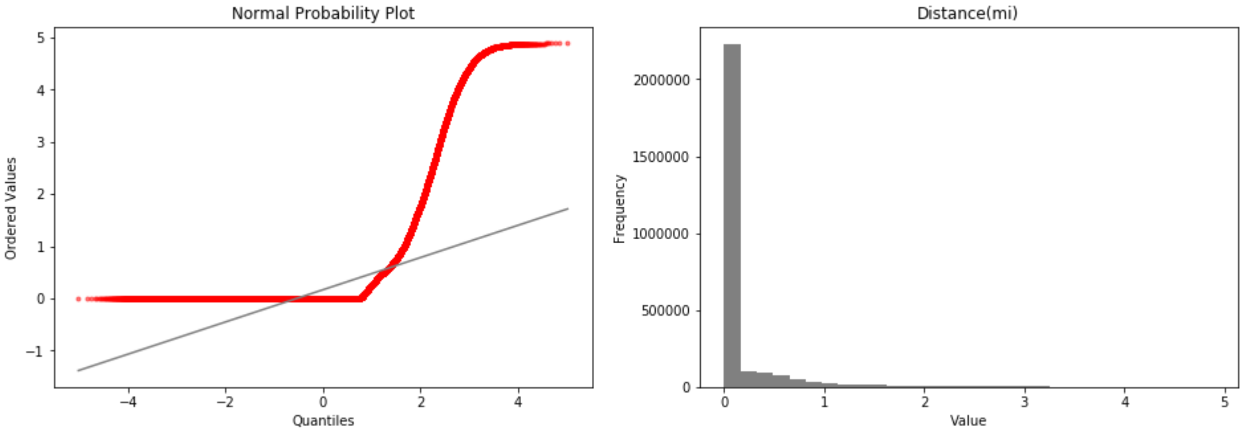


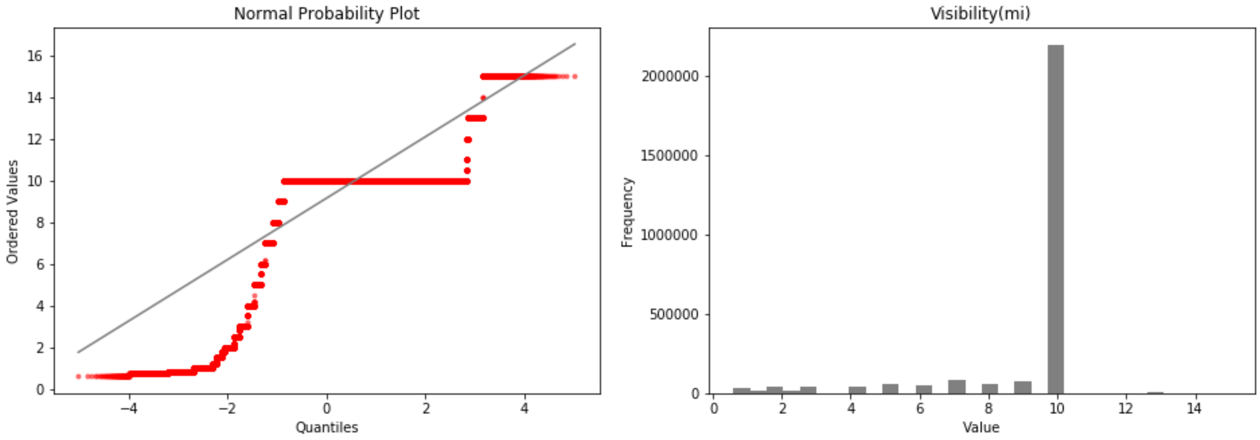


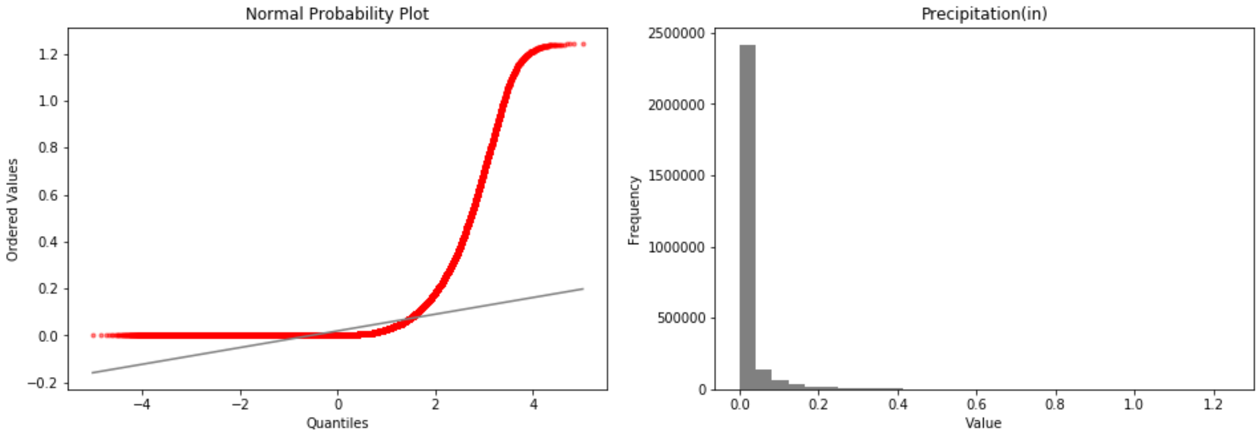




The graphs also highlight several features that have a predominant or possibly default setting.





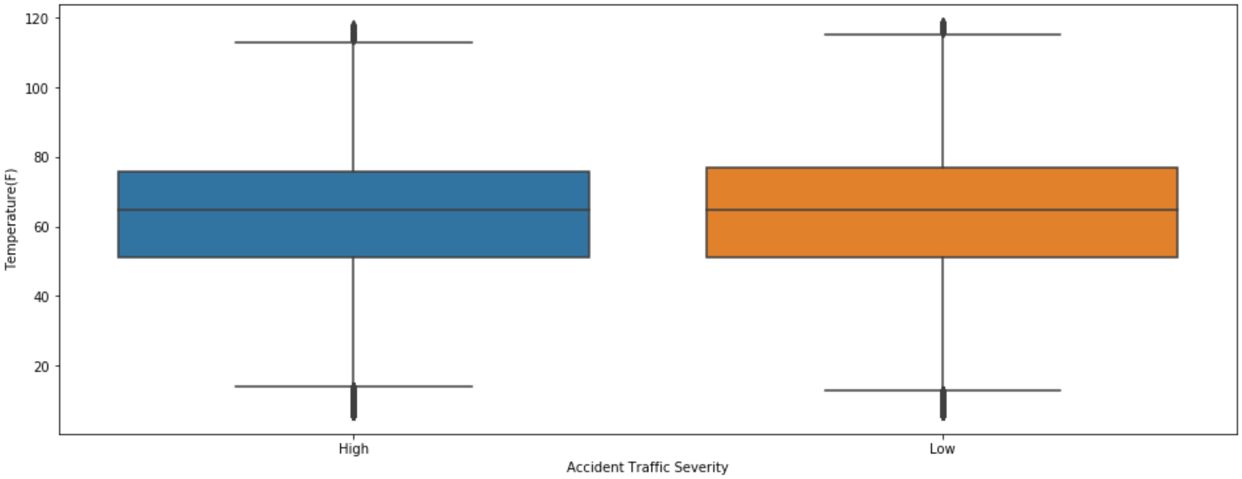


Some feature engineering can restate these metrics in a binary way, such as Default / Not Default.

Difference of Means:

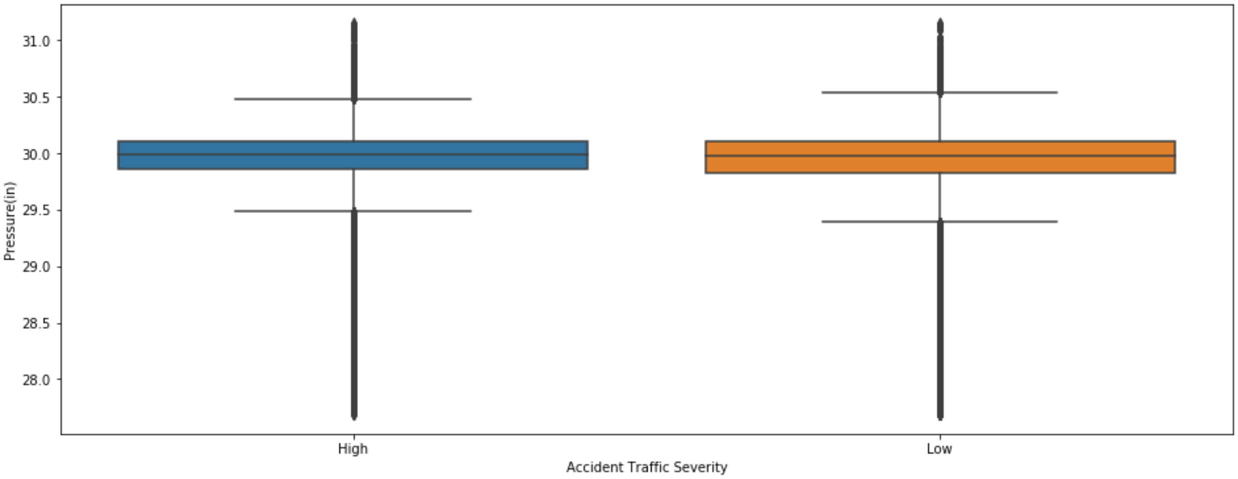
To simplify analysis and prepare for classification, the four point Severity scale is condensed into binary, Low/High or 0/1 - combining the former 1 with 2, and 3 with 4. The analysis will compare the differences between means of the Low and High Severity of traffic after accidents.

All comparisons between Low and High Severity appear visibly equal in the following example box plots but numerically slightly different.

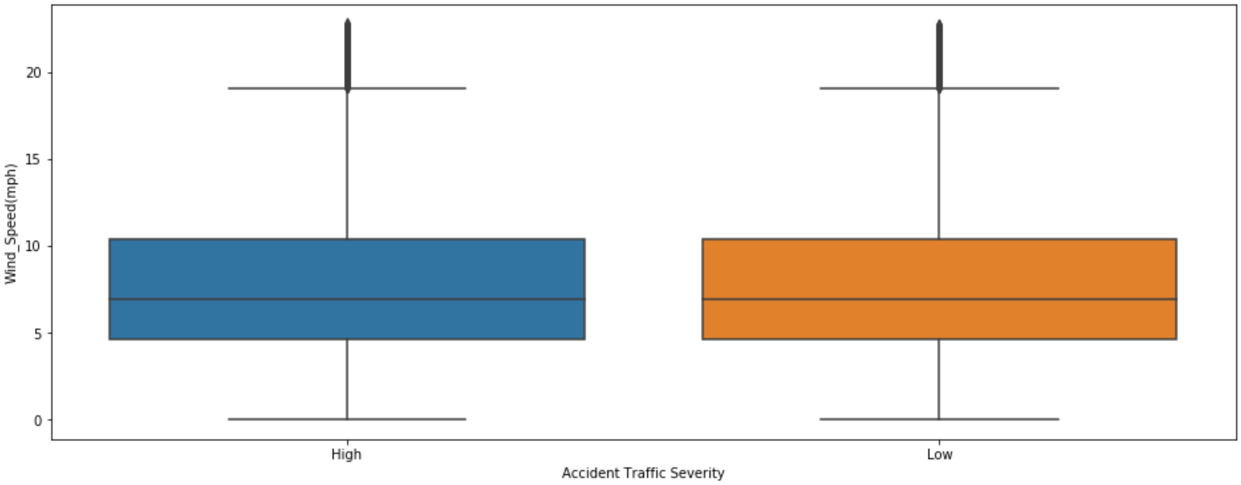


High Severity Temp Mean: 62.631576185300524
Low Severity Temp Mean: 63.253038855079296

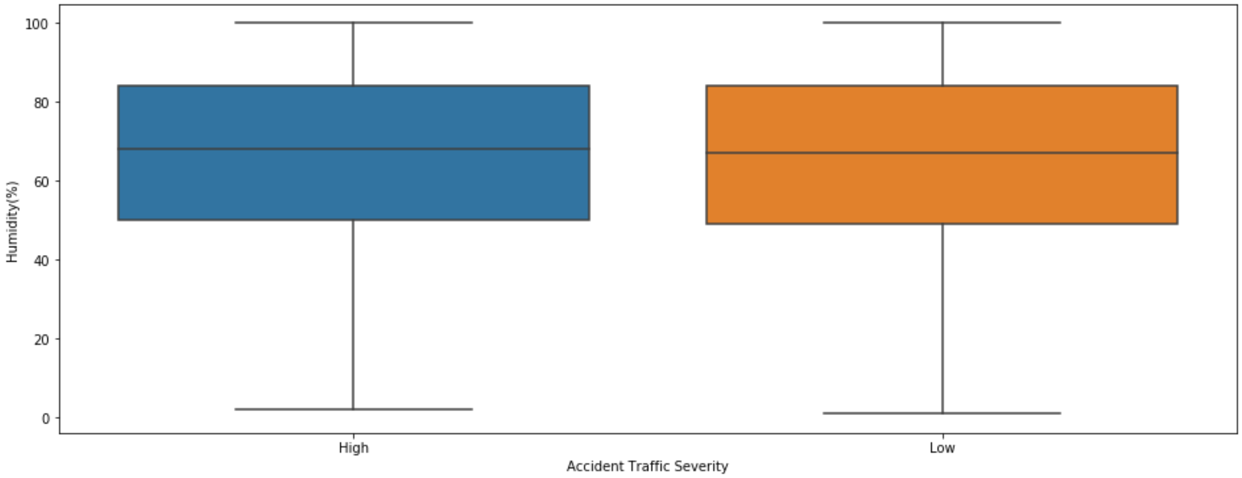
```
Mean Diff: -0.6214626697787722
H0 Diff: 0
```



```
High Severity Pressure Mean: 29.932684963998195
Low Severity Pressure Mean: 29.90275656117535
Mean Diff: 0.029928402822843348
H0 Diff: 0
```



```
High Severity Wind Speed Mean: 7.999216873394668
Low Severity Wind Speed Mean: 7.7370550303059735
Mean Diff: 0.2621618430886947
H0 Diff: 0
```



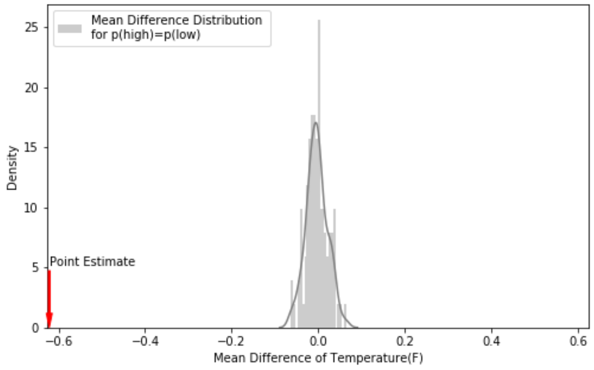
High Severity Humidity Mean: 66.08772904320682
Low Severity Humidity Mean: 65.21196348289683
Mean Diff: 0.8757655603099863
H0 Diff: 0

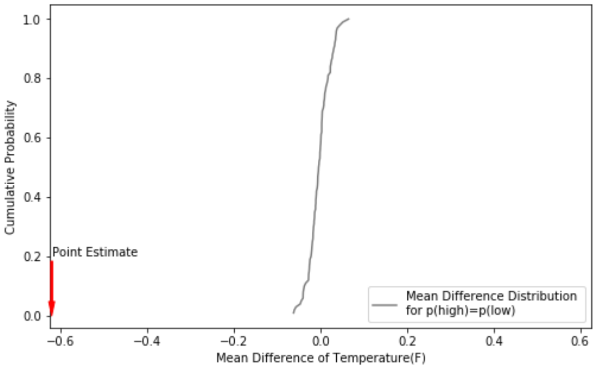
There is some difference between the means of the two distributions, but without a statistical significance test, it can't be sure that this difference isn't simply due to chance of the sample and that the distributions are, in fact, the same.

Null Hypothesis Tests:

A Z-Test is used to better understand whether the distributions are statistically significantly different. The High and Low data are mixed together in many random permutations, divided, and the means of these new random groups distributed.

H0: The distributions between High Severity and Low Severity Accident Traffic are the same.





```
p: 0.0
CI: [-0.05213991  0.04192216]
ME: 0.045188565274005346
```

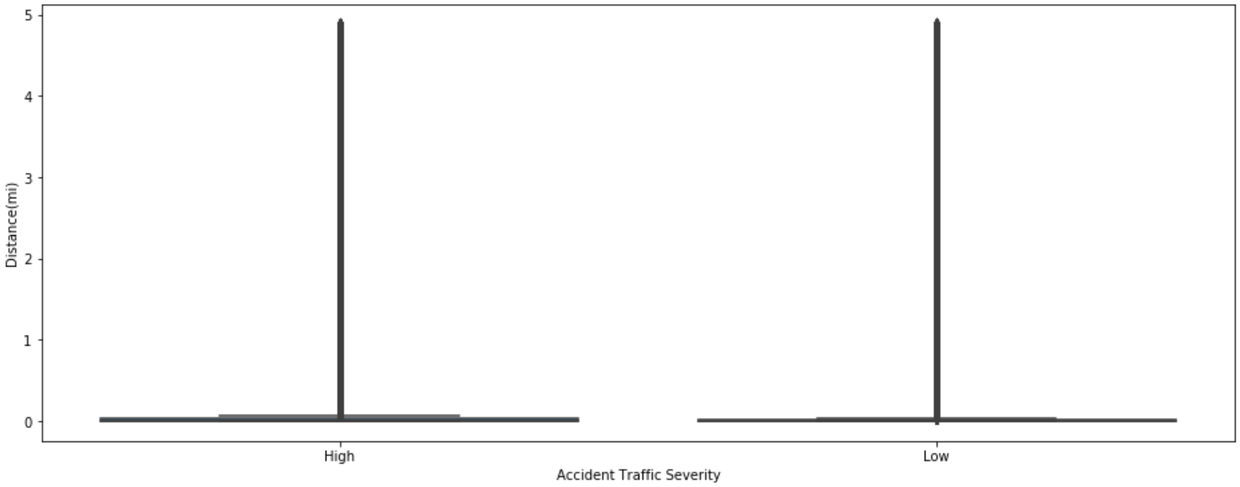
The original mean temperature difference between High and Low groups was -0.6 as identified by the arrow. The Confidence Interval of our random permutations includes zero, no difference, and does not contain -0.6. The Margin of Error is low at 0.045. It is clear that our observed mean difference is statistically significant.

The p-values calculate to zero every time, which would reject the null hypothesis that the means are equal. Additional two-sample t-tests, as well as the nonparametric Mann-Whitney U test also confirmed nearly zero p-values. This is due to our large dataset size, the “curse of the central limit theorem”, and that large datasets are more likely to find statistically significant relationships.

In similar tests, subsamples of 1,000 records were used in 100 comparisons and p-values continued to remain near zero.

One metric to check further, the mean distance of High Severity accident traffic is twice as large as Low Severity. The odd box plot implies outliers affecting both categories:

```
High Severity Distance Mean: 0.2512084166855638
Low Severity Distance Mean: 0.12522722810616832
Mean Diff: 0.1259811885793955
```



Windsorizing was applied to avoid top and bottom 10% of outliers and it retains dissimilar means.

```
High Severity Distance Mean: 0.14201143525515877
Low Severity Distance Mean: 0.06472388298910384
Mean Diff: 0.07728755226605494
```

And, of course, p-value = 0

Due to possible record size, the p-values reject the null hypothesis that the means were equal. Let us move on to machine learning algorithms to identify any trends.

6. Machine Learning Models

As noted earlier, the first model to try will be a logistic regression classifier to predict whether accidents will have a High or Low impact on traffic.

This is different from the statistical tests above that are looking at a hypothesis and decide whether or not the result is very unlikely to have happened by chance, but rather this analysis helps to decide whether or not it could be beneficial to increase model complexity by using a feature due to its predictive power.

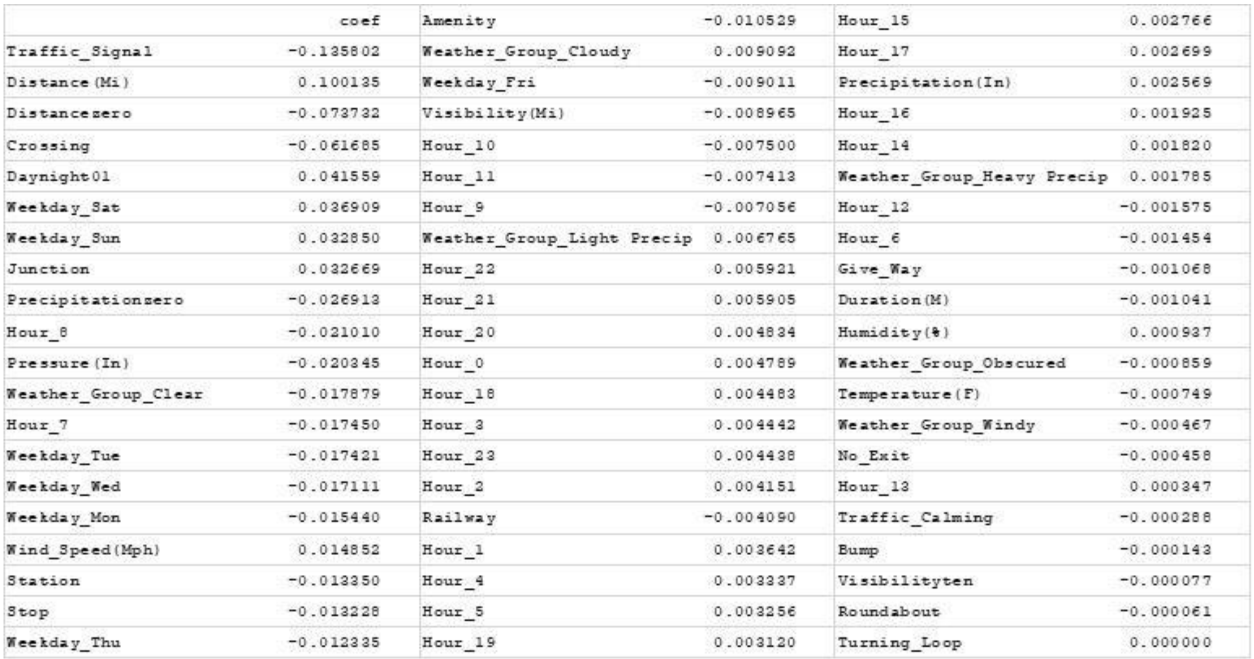
Preprocessing modifications:

- The 1-4 Severity now has Low/High and 0/1 columns
- Distance, Precipitation and Visibility have very low variance and represented as 0/1 for their not default/default values
- Civil Twilight indicating Day/Night becomes 0/1
- Weather Condition has 106 descriptions, they were consolidated into 6 recurring themes:
  clear, cloudy, windy, light precipitation, heavy precipitation, and obscured (fog, dusty, etc.)
- The weekdays, hour of day, and weather groups were broken out into binary columns
- Dropped Wind Chill as it had high correlation to Temperature

Logistic Regression Classifier:

Starting with a Logistic Regression on the 62 features and a standard scalar, the baseline accuracy is 69.4%. Obviously, not all of these features should be considered.

The logistic regression .coef_ attribute is used to rank and print the coefficients used on each feature.



	coef	Amenity	-0.010529	Hour_15	0.002766
Traffic_Signal	-0.135802	Weather_Group_Cloudy	0.009092	Hour_17	0.002699
Distance(Mi)	0.100135	Weekday_Fri	-0.009011	Precipitation(In)	0.002569
Distancezero	-0.073732	Visibility(Mi)	-0.008965	Hour_16	0.001925
Crossing	-0.061685	Hour_10	-0.007500	Hour_14	0.001820
Daynight01	0.041559	Hour_11	-0.007413	Weather_Group_Heavy Precip	0.001785
Weekday_Sat	0.036909	Hour_9	-0.007056	Hour_12	-0.001575
Weekday_Sun	0.032850	Weather_Group_Light Precip	0.006765	Hour_6	-0.001454
Junction	0.032669	Hour_22	0.005921	Give_Way	-0.001068
Precipitationzero	-0.026913	Hour_21	0.005905	Duration(M)	-0.001041
Hour_8	-0.021010	Hour_20	0.004834	Humidity(%)	0.000937
Pressure(In)	-0.020345	Hour_0	0.004789	Weather_Group_Obscured	-0.000859
Weather_Group_Clear	-0.017879	Hour_18	0.004483	Temperature(F)	-0.000749
Hour_7	-0.017450	Hour_3	0.004442	Weather_Group_Windy	-0.000467
Weekday_Tue	-0.017421	Hour_23	0.004438	No_Exit	-0.000458
Weekday_Wed	-0.017111	Hour_2	0.004151	Hour_13	0.000347
Weekday_Mon	-0.015440	Railway	-0.004090	Traffic_Calming	-0.000288
Wind_Speed(Mph)	0.014852	Hour_1	0.003642	Bump	-0.000143
Station	-0.013350	Hour_4	0.003337	Visibilityten	-0.000077
Stop	-0.013228	Hour_5	0.003256	Roundabout	-0.000061
Weekday_Thu	-0.012335	Hour_19	0.003120	Turning_Loop	0.000000

An interesting ranking. Weather is not well represented at the top. Traffic signals, intersections and crosswalks are far more indicative of traffic severity than weather. One of the created binary fields, Precipitation01, made the top ten. Binary Distance is redundant as Distance(mi) is ranked second. The weekend ranked well, identified as much less frequent accident days during the EDA, as well as the importance of commute times beginning at 7-8 am. Pressure can indicate good weather when it increases and storms when it falls. I am surprised to see Heavy Precipitation ranking so low, maybe due to causing slower travel speeds.

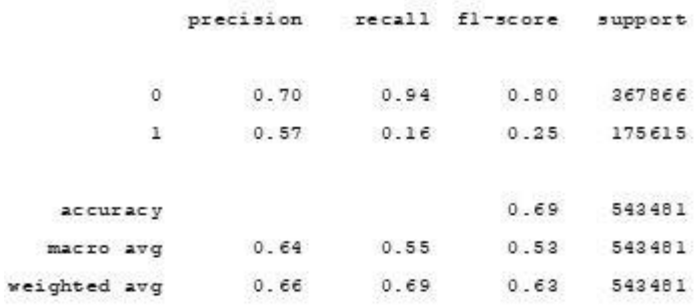
The line will be drawn at coefficients of at least +/- 0.01 away from zero.

The second logistic regression uses 1/3rd of the features and scores 69.0% accuracy, down slightly more than 1/3rd of a percent. We continue to evaluate its performance.

Cross-fold validation yields five similar scores, indicating the model is not overfitting:

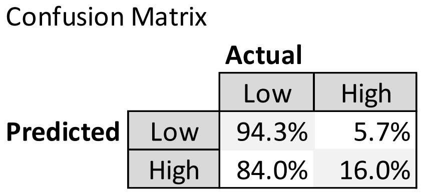
0.689   0.693   0.689   0.685   0.685
Average 5-Fold CV Score: 0.688

The classification report shows a very concerning figure:



```
              precision    recall  f1-score   support

           0       0.70      0.94      0.80    367866
           1       0.57      0.16      0.25    175615

    accuracy                           0.69    543481
   macro avg       0.64      0.55      0.53    543481
weighted avg       0.66      0.69      0.63    543481
```

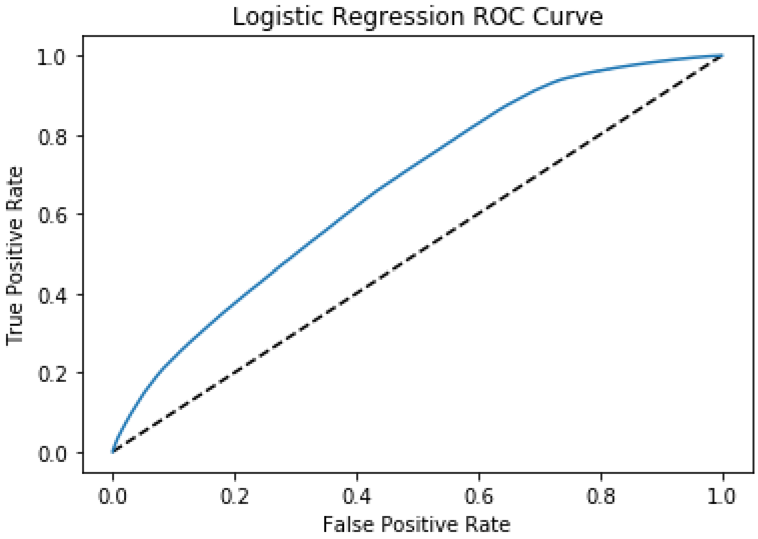
Although recall is high for the Low Severity accidents, the model only identifies 16% of the High Severity accidents.



Confusion Matrix

		Actual	
		Low	High
Predicted	Low	94.3%	5.7%
	High	84.0%	16.0%

This confusion matrix adds that 84% are false negatives, meaning they were actually high severity accident traffic, but incorrectly identified as low severity. This is a substantial problem as motorists would be led erroneously into waiting in extensive traffic delays. The model's focus must be to reduce false negatives.



The ROC score of 66.8% and graph also shows the classifier has poor discriminative ability.

Another model may improve the outcome.

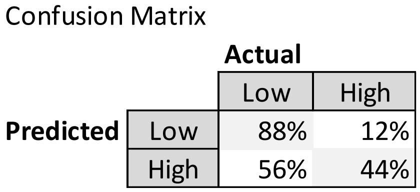
Random Forest Classifier:

This model will use the dummy variables created for day, hour and weather group. Per the robustness of Random Forest classifiers, I have skipped two pre-processing steps:

- Did not remove Wind Chill due to high correlation to Temperature
- Did not create binary versions of low variance features

The RF classifier almost added four percentage points over Logistic Regression:
RF Accuracy: 0.7395180328291145

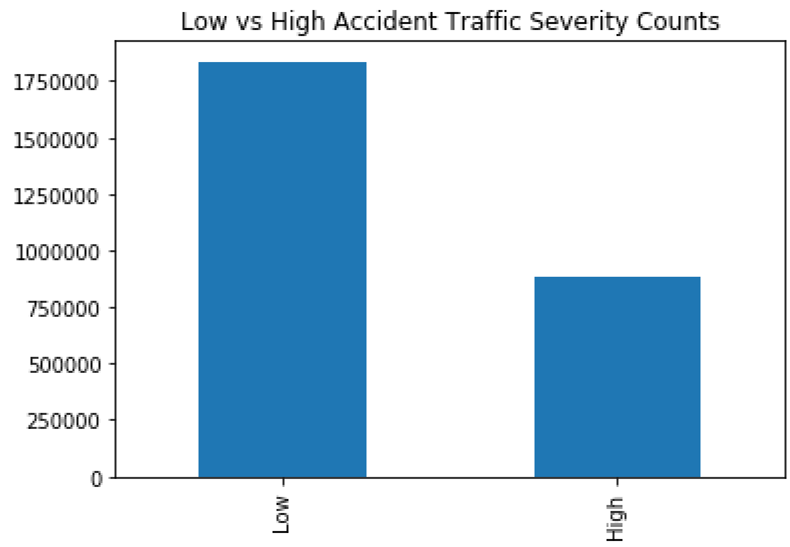
And there was progress at lowering the false negatives, the concern with LR. True negatives more than doubled, and a 30% reduction of false negatives:



Confusion Matrix

		Actual	
		Low	High
Predicted	Low	88%	12%
	High	56%	44%

As a next step, check if High Severity accident traffic is under-represented:



There is a 2:1 difference that may be impacting the model's sensitivity. There is a lot of data, so instead of creating copies of High Severity records to match the Low quantity, the Low quantity was randomly reduced to match the count of High Severity. 880K of each:

```
1    880,347
0    880,347
```

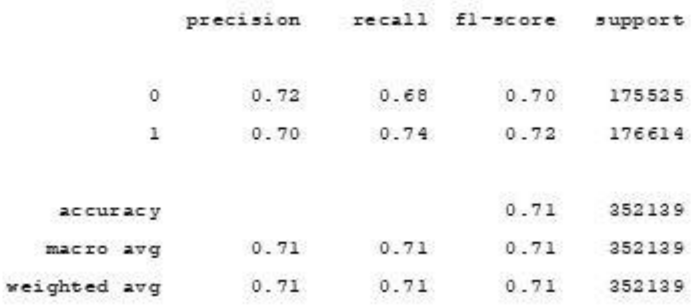
Another run of the model produces:
RF Accuracy: 0.7081237806661574

Confusion Matrix

		Actual	
		Low	High
Predicted	Low	68%	32%
	High	26%	74%

False Positives are cut by 50% further, and a higher rate of identifying High Severity Accident Traffic than Low. As noted previously, the model needs to minimize false positives, which is High Severity incorrectly identified as Low Severity. This improvement is worth the tradeoff of reducing overall accuracy by three percentage points.

If a motorist is told an accident will cause High Severity traffic, in error, they will take an alternate route, or be pleasantly surprised when actual traffic is less than expected, and thus a much less problematic prediction than the opposite.

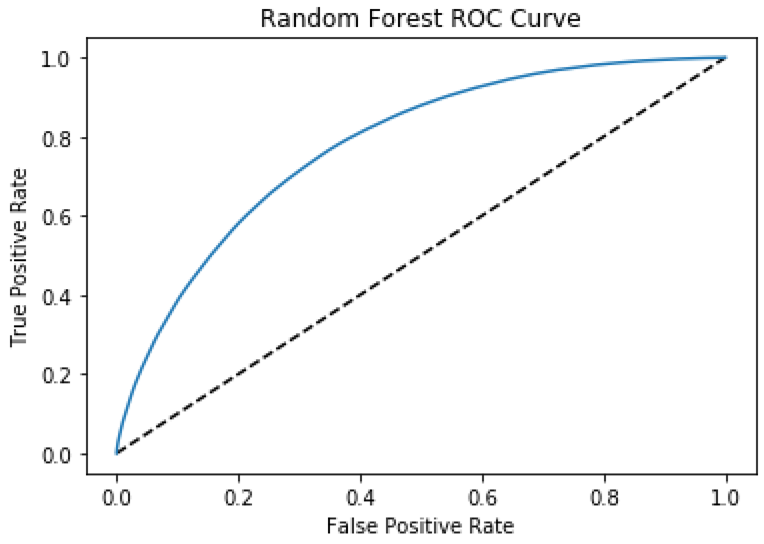


```
              precision    recall  f1-score   support

           0       0.72      0.68      0.70    175525
           1       0.70      0.74      0.72    176614

    accuracy                           0.71    352139
   macro avg       0.71      0.71      0.71    352139
weighted avg       0.71      0.71      0.71    352139
```

The weighted average F1 score increased from 64% to 71% and High Severity Recall has improved from 16% to 74%.

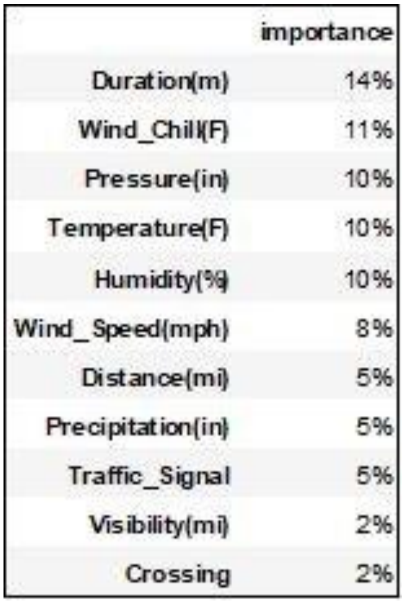
Five-fold cross-validation continues to indicate no over-fitting:

0.585 0.627 0.556 0.628 0.609
Average 5-Fold CV Score: 0.601



And the area under curve improved from 66.8 to 77.9

<u>Feature Importance</u>:

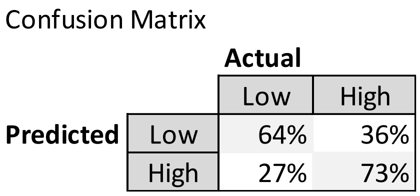


	importance
Duration(m)	14%
Wind_Chill(F)	11%
Pressure(in)	10%
Temperature(F)	10%
Humidity(%)	10%
Wind_Speed(mph)	8%
Distance(mi)	5%
Precipitation(in)	5%
Traffic_Signal	5%
Visibility(mi)	2%
Crossing	2%

The beginning of the Random Forest feature ranking relied considerably more upon weather features than Logistic Regression, in nine of the first ten ranks. Traffic Signal and Crosswalk are again the top non-weather features, followed by all of the days close together.

Is there more improvement ignoring days, hours, weather groups and everything lower?

RF Accuracy: 0.686714621214918



Confusion Matrix

		Actual	
		Low	High
Predicted	Low	64%	36%
	High	27%	73%

Slightly underperforming what we had previously with all the fields. False Positives only went up 1 point, but reduced from 60 fields to 11 to lesson computational load for hyperparameter tuning.

Hyperparameter Tuning:

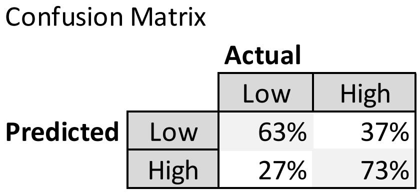
A randomized search was used on several parameters, fitting 3 folds for each of 50 candidates, totalling 150 fits:

- n_estimators = [int(x) for x in np.linspace(start = 100, stop = 1000, num = 10)]
- max_features = ['auto', 'sqrt']
- max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
- min_samples_split = [2, 5, 10]
- min_samples_leaf = [1, 2, 4]

Eighty minutes to run, results were:

```
{'n_estimators': 200,
 'min_samples_split': 2,
 'min_samples_leaf': 4,
 'max_features': 'auto',
 'max_depth': 20}
```

RF Accuracy: 0.6810577641215543



Confusion Matrix

		Actual	
		Low	High
Predicted	Low	63%	37%
	High	27%	73%

The random tuning declined our model performance slightly, reducing half a percentage on accuracy and a one point shift from true positives to false positives. With more time, a more thorough grid search may find slight improvement from optimal tuning.

7. Conclusion

After a traffic accident occurs, it is feasible to apply weather conditions to differentiate whether resulting traffic will be severe or not. The added information would be beneficial to motorists to update their travel plans. Such warnings could be sent as alert messages to smartphones and GPS devices, not only by travel applications but also from weather applications which utilize users' location.

This novice exercise with binary classifications showed a potential with nearly 75% accuracy identifying high severity traffic. Other predictive machine learning tools and additional information on roadways, i.e. number of lanes, average speed, volume, etc., should generate more consistent notifications.