Classification of Vehicle Crash Severity based on Weather Conditions

Christopher Gardner

1 Introduction

1.1 Background

Vehicle Crashes are a leading cause of death in the US, causing approximately 35,000 fatalities in 2015.[1] This does not include the number of incidents which do not cause death, which bumps up to almost 4.4 million individuals who were injured enough to require medical attention. With so many incidents it is necessary to be able to respond quickly to avoid serious medical injuries. The average response time to a vehicle collision in the US is approximately 9 minutes with longer response time being significantly associated with higher mortality rates.[2] Thus it is vital that the appropriate resources are available when such incidents occur.

1.2 Interest

The primary party we are interested in informing with this study are healthcare systems which would be responsible for having emergency medical technicians (EMTs) available to respond to these incidents. This would ideally serve a two-fold purpose of reducing deaths associated with vehicle collisions as well as provide a more efficient budget for healthcare systems. Being able to also triage potential accidents into their high severity and low severity cases would further benefit this resource allocation strategy.

2 Data

2.1 Data Sources

We have gathered two separate data sources. The one we will be discussing for the remainder of the study is from GISWEB Collision data. We have also done a similar/preliminary data analysis by using a combined data set with NY vehicle collisions gathered from data.gov and weather conditions from NOAA. This previous study had very little correlation/predictive power, so we used the GISWEB data which has more relevant data features without cleaning. The GISWEB data has records of vehicle collisions alongside weather conditions, road conditions, driver awareness, location and other aspects of collisions. We can extract these features to gain insight into vehicle collision causes and create a model to predict high severity.

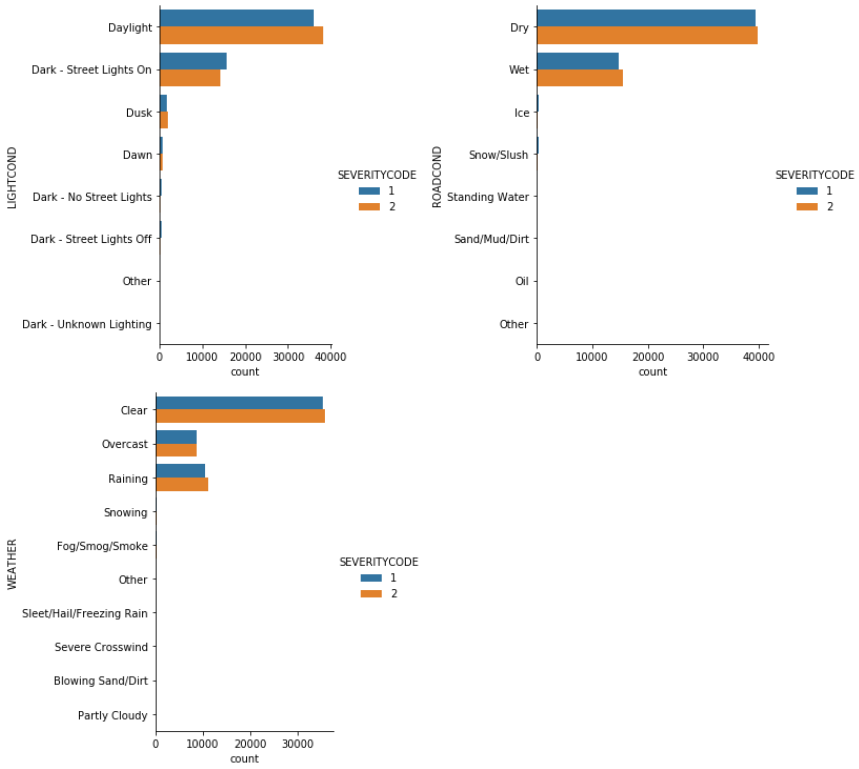
2.2 Data Cleaning

The data that we have gathered is presented in a mostly clean form, with relevant features already created and severity levels already defined. Because of this, our data cleaning involves removing unnecessary columns and dealing with missing/unknown values. For dropping columns we need to remove information that is encoded within the SEVERITYCODE feature such as number of injuries and serious injuries. Other information that is not relevant to our model/analysis such as location, date, vehicle count, and driver attention among others are also dropped. We are primarily concerned with features which could impact collisions and that we can understand beforehand to predict crash severity before it occurs.

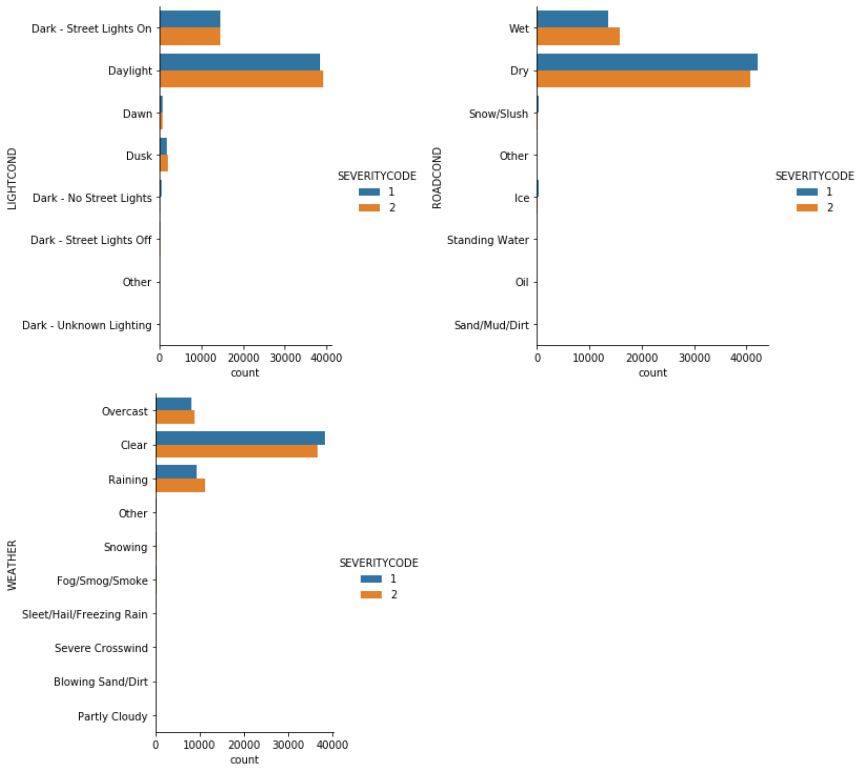
The second part of data cleaning was dealing with unknown data regarding weather and road conditions. In this case we have two options. The first is to drop the data rows that are unknown, the primary motivator for this is that even if it is a good predictor it is a meaningless value. Even if it were good for prediction the insight would have no value to the user. EX: Unknown weather is safer than Clear Sky + Dry roads. The other line of thought would be to replace the data value unknown with the most common (mode) value. This might work well given the assumption that conditions were normal, so the data collector/report did not include details about the weather. We will explore both options.

3 Methodology / Data Exploration and Preparation

For the exploratory data analysis we made visualizations to show the distribution between categorical values compared to the severity code. Knowing that unknown values need to either be dropped or replaced we visualize data features we are interested in such as ROADCOND, LIGHTCOND, and WEATHER.



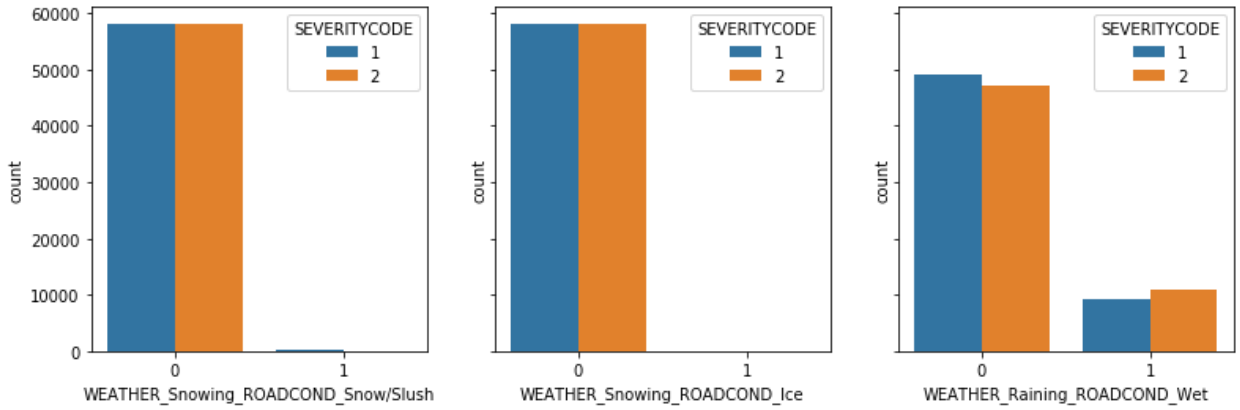
The figures above show that when Unknowns are removed the distributions between severitycode levels is very slim. Next we'll look at the bar charts when we replace unknown with the most common data values.



When replacing the unknown values you can see that compared to dropping the unknown values keeping them provides a lot of valuable prediction power. Since both options had valid reasoning we will move forward with unknown values replaced with the mode.

Next we needed to do data preparation to allow the data to be passed into our predictive model. We parsed LIGHTCOND down to two values "daylight" and "dark" and used one-hot encoding to create numerical features out of the categorical columns we chose (ADDRTYPE, LIGHTCOND, WEATHERCOND, ROADCOND)

After that we decided to bolster our data by adding feature interactions to our dataset. We do this by multiplying the two features we want to see interactions for. We decided to do this to combine weather and road conditions.



As you can see above, some of the interactions create very useful features that we can apply to the model.

4 Results

After completing data preparation and feature creating, we tested 3 different classification models: LogisticRegression, DecisionTree, and K-NN. We use F-1 Score, Precision and Recall on the high-severity class which we denote as “Injury” to compare these models. Since we would also like to improve the models, we implemented a grid search to find the best parameters.

Model	Precision	Recall	F1-Score
Logistic Regression	0.64	0.48	0.55
Decision Tree	0.64	0.47	0.54
K-nn	0.60	0.48	0.53

For these prediction models they are all very similar, but the Logistic Regression classifier performs the best in both precision and recall, and therefore F1-score because it is a combination of the two. The metrics in the table above show us that 64% of the samples classified as “Injury” were in fact injuries, but we only capture 48% of the total possible “Injury” classes. To attempt improve our model even further we became more specific with our prediction criteria, only accepting absolute correlation greater than 0.2. Below the new Logistic Regression evaluation metrics are show. The model did not see improvement with this increased selectivity.

Model	Precision	Recall	F1-Score
Logistic Regression	0.64	0.48	0.55

Along with the model we were able to find some high correlation features which give us some insights into what features to investigate for future modelling and data collection. Extra attention can be given toward feature interactions between the final prediction features of our model listed below.

Features
Alley
Block
Intersection

WEATHER_Clear
WEATHER_Raining
ROADCOND_Dry
ROADCOND_Wet
ROADCOND_Snow/Slush
WEATHER_Clear_ROADCOND_Dry

5 Discussion

After completing the study of data exploration and modelling we have a few recommendations for health systems interested in optimizing EMT shift scheduling for fast emergency response times and to reduce costs due to overscheduling.

The first recommendation is that the address type has the highest impact on collision severity. With Injuries being much more correlated with intersections than blocks or alleys. This means that positioning emergency services in locations with more intersections would allow for quicker response times in those situations which are more likely to result in injuries.

The second insight is that weather and road conditions can help predict crash severity and such shift scheduling for emergency response times could be elevated on rainy days where you expect wet road conditions. However fewer emergency staff may be needed on clear weather and dry road days.

Along with these insights we have many new data and features to explore in the future for improving our predictions. For future studies we would like to expand our dataset to look at location (for high severity hotspots) and time of crash (is there a more dangerous time to drive. Along with these new data features to gain insight on we would also like to explore collision frequency as well. By exploring collision frequency alongside crash severity we could potentially mark high risk days which coincide with high frequency days. Having enough emergency technicians available on those days will be vital to reducing severe injuries and fatalities.

6 Conclusion

This study has provided us with a model to capture most high severity incidents, which will allow health systems to better allocate resources when high severity incidents are more likely. We have also identified new features and questions we would like to answer in future studies to expand what was started in this study.

7 Resources

[1] https://www.cdc.gov/injury/wisqars/overview/key_data.html#:~:text=Motor%20vehicle%20crashes%20are%20a,prescription%20opioid%20overdoses%20in%202015.

[2]
https://pubmed.ncbi.nlm.nih.gov/30725080/#:~:text=Abstract,to%20influence%20trauma%20patient%20survival.&text=The%20median%20county%20response%20time,%2C%207%2D11)%20minutes.