

predicting the probability of severe accidents based on weather conditions

By Andi Frkovich



September 10, 2020

**Part 1: Introduction.**

Seattle Police Department (SPD) is responsible for responding to and clearing traffic accidents throughout the Seattle Metropolitan area and coordinating any necessary medical care. However, the Seattle Police Department is also responsible for responding to a variety of other local crises and incidents and must decide how to allocate manpower accordingly. Predicted weather conditions on different days and seasons may help police predict the probability of severe traffic accidents throughout the city. Police could then use this information to optimize assigned shifts and manpower allocation between different SPD units (i.e., it could be used to predict how many police should be assigned to traffic units on a particular day).

Data that would help us understand the impact of weather on traffic accident severity includes meteorological weather, road conditions, and light conditions. This project hopes to use these data to ascertain whether and to what extent weather conditions can be used to predict traffic accident severity.

SPD would be interested in predicting the probability of severe traffic accidents so that they can better allocate departmental manpower and resources to respond to such accidents. Others who may be interested in this data may be local hospitals, which can also use the information to optimize staffing and equipment availability, and local news and public health organizations, which could leverage the information to issue timely health warnings to the public.

**Part II: Data Acquisition Cleaning, & Exploration**

Data on weather conditions and the severity of co-occurring traffic incidents can be found in a Kaggle dataset located ​[here](http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0.csv)​. The original dataset had 221,226 features. Because the data set has so many features, we can safely drop the features with missing weather data; the resulting sample still has 194,310 attributes, a very robust n sample.

There are numerous problems with these data. First, some of the attributes had missing data. Because the dataset was so large, I decided to drop the missing data from the set. Even after dropping the missing data, there were still over 190,000 features, which left me with a more than adequate dataset. Second, various attributes were over-weighted to certain outcomes. For example, ‘Light conditions’ was over-weighted to ‘daylight’, ‘Road Conditions’ was overweighted to ‘dry’, and ‘Weather’ was ‘overweighted’ to ‘clear’. To fix the issue, I decreased the weighting of those particular outcomes and normalized the dataset. I did not discover a significant number of outliers. Finally, in order to make the data work in predictive models, I converted descriptive data (i.e., ‘dry road’) to numeric data where the number assigned corresponds to the severity of conditions.

The ‘severity code’ attribute, which incorporates numerous other attributes within the dataset to holistically estimate accident severity, was determined to be the target variable. The numerous features that feed into the ‘severity code’ attribute were not kept, as they are already reflected in the resultant ‘severity code’.

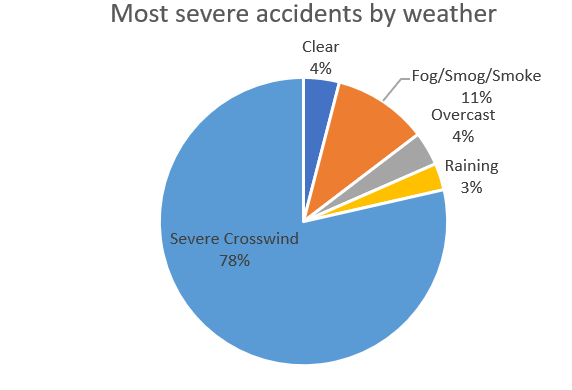
The data contained numerous administrative attributes, which would solely be used to help police keep administrative track of the data. Because these attributes are arbitrary administrative descriptors, these attributes were also dropped. Location data was dropped, as it does not directly contribute to our problem statement.

The remaining attributes are weather conditions, road conditions, and light conditions. These weather attributes will be used to predict the target variable of ‘accident severity’. After data cleaning and features selection, there are 194,310 features and four attributes.

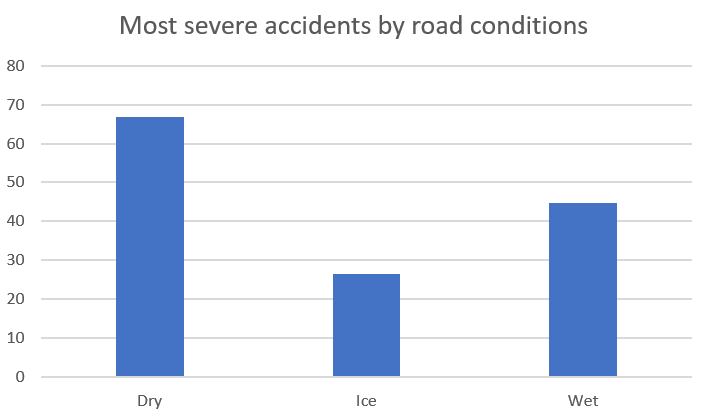
Severity of car accidents (‘Severity code’) was an attribute in the dataset, and encapsulated many variables (number of people involved, damage to car, etc.) that could be used as a proxy for accident severity. Thus, ‘severity code’ was used as the target variable.

Initial exploration of the data revealed significant relationships between weather conditions and accident severity.

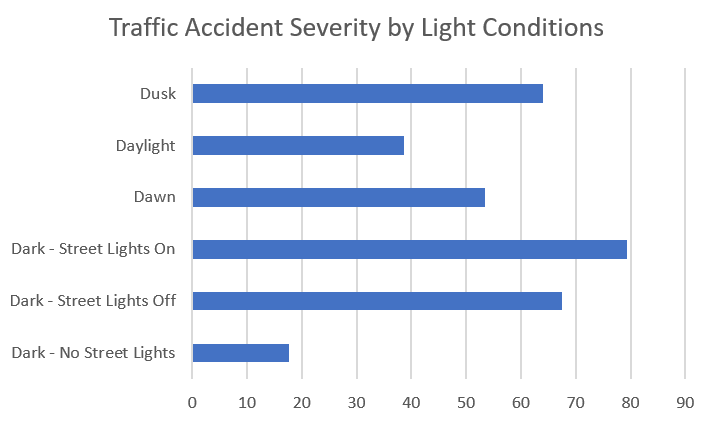
In terms of weather, strong crosswinds were by far the greatest predictor of the most severe accidents, with 78% of category three accidents occurring during periods of strong headwinds (after re-weighting data due to oversampling of clear conditions).



Poor road conditions seemed to cause more accidents, but didn’t result in the most severe accidents, probably due to slower speeds and increased driver caution. Interestingly, the most accidents occurred in dry road conditions.



There was also a demonstrable relationship between light conditions and traffic accident severity. After re-weighting the data, we can see that the vast majority of the most severe accidents (category 3) occur in reduced light conditions.

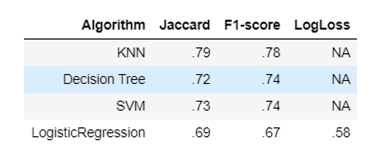


**Part IV: Methodology**

Based on course learnings, a variety of models can be applied to understand how weather affects the probability of experiencing a severe traffic accident. Because this particular study aims to predict the probability of engaging in a severe (‘Category 3’) accident based on weather conditions- specifically, meteorological weather, road conditions, and light conditions- we focus primarily on classification models. Specifically, we apply the K-Nearest Neighbor, Decision tree, Logistic Regression, and SVM models. Before applying these models, all categorical values of data had to be converted to numeric values and numerous attributes had to be re-weighted to account for overrepresentation. The Jaccard and F-1 Scores were used as the primary evaluation metrics.

**Part IV: Results**:

To apply the classification models, I divided the data into two classes (Severity Code>=0 but <3, and severity code of >=3). I then re-weighted data classes to account for over-representation. Jaccard and F-1 Scores were the primary measures of accuracy, though I also calculated LogLoss for the logistic regression model. The KNN model, with K=8, proved to be the most accurate model, although the accuracy of the models was relatively comparable. The KNN model yielded a Jaccard score of .79 and an F-1 score of .78.



**Part V: Conclusions**.

In this project, I evaluated the relationship between weather conditions and the probability of severe car accidents. I identified severity code as the target variable. Weather, road conditions, and light conditions were the independent variables. Because this study is primarily concerned with the probability of a severe accident, I applied a variety of classification models to predict such probability. Findings reveal that driving certain severe weather conditions, i.e., strong winds and low-light conditions, likely increases the probability of being involved in a severe car accident. These models will aid Seattle Police Department and surrounding hospitals with predictive resource allocation.

**Part VI: future directions**.

In this project, I was able to achieve ~78% accuracy in classification. However, significant variance still remains. The models could be improved with more granular (and more accurately scaled) weather data. For example, instead of collecting descriptive data for road conditions, data on the relative slipperiness of roads could be a more accurate indicator. More reliably scaled information on light conditions would also help improve the accuracy of the models (for example, is a street without streetlight or a street with its streetlights off darker?). More data, specifically on how weather impacts road and light conditions, would be helpful. For local hospitals, it could also be helpful to know if certain weather conditions are more heavily linked to specific injury severity or types. These data could help significantly improve the model.