Predicting Accident Probability

David Cavanaugh

Data Science Capstone Project

9/18/20 8:45:00 AM

TAble of  
Contents

Abstract 1

Introduction 1

Problem 2

Interest 2

Methodology 3

Data Sources 3

Data Cleaning 3

Feature Selection 6

Results 8

Discussion 10

Conclusion 10

# Abstract

# Introduction

This case study will be used to develop a system to predict the severity of an accident. The system will warn the user, given the weather and the road conditions about the possibility of getting into a car accident and how severe it would be. The desired outcome is to warn the user so they would drive more carefully or even change their travel route.

## Problem

Say you are driving to another city for work or to visit some friends. It is rainy and windy. On the way to your destination, you come across a terrible traffic jam on the other side of the highway. Long lines of cars are barely moving. As you keep driving, police cars start appearing from afar, shutting down the highway. There is an accident and a helicopter is transporting people involved in the crash to the nearest hospital. The victims must be in critical condition for all of this to be happening.

## Interest

Now, wouldn't it be great if there is something in place that could warn you, given the weather and the road conditions, about the possibility of you getting into a car accident and how severe it would be. The advance warning could prompt you to drive more carefully or even change your travel plans if you are able to.

# Methodology

## Data Sources

The data used to train and evaluate the model is the collision data set from the SDOT Traffic Management Division, Traffic Records Group. The data set is updated weekly from 2004 to the present. The data set is compiled from all collisions provided by the Seattle Police department and recorded by the Traffic Records Group. Collisions will display at the intersection or mid-block of a segment.

The current collision data may be acquired from [http://data-seattlecitygis.opendata.arcgis.com](http://data-seattlecitygis.opendata.arcgis.com/).

The collision data attribute information may be acquired from <https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf>.

A copy of the collision data (15 September 2020) was used to generate the results for this report.

## Data Cleaning

An initial review of the dataset indicates that a number of features that may be safely eliminated. These features are used for various bookkeeping functions or are textual descriptions of categorical data.

* X, Y: Coordinates of a collision. Redundant with LOCATION
* COLDETKEY: bookkeeping
* REPORTNO: bookkeeping
* STATUS: bookkeeping
* INTKEY: bookkeeping
* EXCEPTRSNCODE, EXCEPTRSNDESC: Not required, exception information is scattered across other features
* INCDATE, INCDTTM: not enough entries to be of use
* SDOT\_COLDESC: redundant with SDOT\_COLCODE
* ST\_COLCODE, ST\_COLDESC: redundant with SDOT\_COLCODE and SDOT\_COLDESC

There are problems with the dataset. There are numerous missing values that need to be filled in.

The *ADDRTYPE*, *WEATHER*, *LIGHTCOND*, *ROADCOND*, and *JUNCTIONTYPE* features all consist of enumerated values. There are a significant number on blank values in these fields. The blank fields will be set to the value of *UNKNOWN* in order to generate a frequency table easier.

The *INATTENTIONIND*, *UNDERINFL*, *PEDROWNOTGRNT*, *SPEEDING*, and *HITPARKEDCAR* are binary values representing either *Yes* or *No*. Blank values were assumed to represent a *No* value. A *1* is assumed to be a *Yes* value while a *0* is assumed to be a *No* value.

The *WEATHER* feather has several outliers. I changed the *WEATHER* categories of *Snowing, Fog/Smog/Snow, Sleet/Hail/Freezing Rain, Blowing Sand/Dirt, Severe Crosswind, Partly Cloudy*, and *Blowing Snow* to *Other*. These categories are not major factors in the data set and can be safely combined. Timestamps are not available. If Timestamps were available, the *Unknown* values could be set to reflect the appropriate weather conditions.

Table 1: Weather Adjustments

| Value | Before | After |
| --- | --- | --- |
| Clear | 114342 | 114342 |
| Unknown | 41724 | 41724 |
| Raining | 34019 | 34019 |
| Overcast | 28504 | 28504 |
| Snowing | 919 |  |
| Other | 851 | 2555 |
| Fog/Smog/Smoke | 577 |  |
| Sleet/Hail/Freezing Rain | 116 |  |
| Blowing Sand/Dirt | 56 |  |
| Severe Crosswind | 26 |  |
| Partly Cloudy | 9 |  |
| Blowing Snow | 1 |  |

The *LIGHTCOND* feather has several outliers. I changed the *LIGHTCOND* categories of *Dark - Street Lights On, Dark - No Street Lights, Dark - Street Lights Off,* and *Dark - Unknown Lighting* to *Dark*. These categories are not major factors in the data set and can be safely combined. Timestamps are not available. If Timestamps were available, the *Unknown* values could be set to reflect the appropriate light conditions.

Table 2: Light Conditions Adjustments

| Value | Before | After |
| --- | --- | --- |
| Daylight | 119149 | 119149 |
| Dark - Street Lights On | 50048 |  |
| Unknown | 40201 | 40201 |
| Dusk | 6074 | 6074 |
| Dawn | 2599 | 2599 |
| Dark - No Street Lights | 1573 |  |
| Dark - Street Lights Off | 1236 |  |
| Other | 244 | 244 |
| Dark - Unknown Lighting | 20 |  |
| Dark |  | 52877 |

The *ROADCOND* feature has several outliers. I changed the *ROADCOND* categories of *Ice, Snow/Slush, Standing Water, Sand/Mud/Dirt*, and *Oil* to *Other*. These categories are not major factors in the data set and can be safely combined.

Table 3: Road Conditions Adjustments

| Value | Before | After |
| --- | --- | --- |
| Dry | 128150 | 128150 |
| Wet | 48711 | 48711 |
| Unknown | 41642 | 41642 |
| Ice | 1231 |  |
| Snow/Slush | 1014 |  |
| Other | 136 | 2641 |
| Standing Water | 119 |  |
| Sand/Mud/Dirt | 77 |  |
| Oil | 64 |  |

The categorical feature values will be converted into integer types to allow a heat map to be generated to help drive feature selection.

## Feature Selection

A heatmap was generated using the 23 available features.

A screen shot of a building

Description automatically generated

Figure 1: Heatmap of Features

The correlation results were then sorted for *SEVERITYCODE\_CAT* target. Correlation results greater than 0.35 were considered relevant.

Table 4: Feature Correlation

| Feature | Correlation |
| --- | --- |
| SEVERITYCODE\_CAT | 1.000000 |
| INJURIES | 0.700338 |
| WEATHER\_CAT | 0.434211 |
| VEHCOUNT | 0.385334 |
| LIGHTCOND\_CAT | 0.380792 |
| PERSONCOUNT | 0.370575 |

The problem statement indicates that the initial feature selection should include the weather and road conditions. The features will be:

* INJURIES
* WEATHER
* VEHCOUNT
* LIGHTCOND
* PERSONCOUNT

The target for the model is *SEVERITYCODE*.

# Results

The Features data frame was then generated using the features listed above. One Hot encoding was done for the *WEATHER* and *LIGHTCOND* features. The Features data frame was then run through the standard scaler to normalize the data frame. Finally, the Features data frame was split into a training data frame and a test data frame.

The following models were used to evaluate the data set:

* K Nearest Neighbor
* Decision Tree
* Support Vector Machine
* Logistic Regression

After running several iterations of the models, it was found that the *FATALITIES* and *SERIOUSINJURIES* features needed to be added to the Features data frame. The following warning message was being generated:

*F-score is ill-defined and being set to 0.0 in labels with no predicted samples.*

Taking the difference between the sets *y\_test* and *yhat* showed that the *SEVERITYCODE* values *Serious Injury Collision* and *Fatality Collision* were not being predicted. Adding the *FATALITIES* and *SERIOUSINJURIES* features resolved the warning.

The following results were observed from the models:

Table 5: Model Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Error | Jaccard Index | F1 Score | Log Loss |
| KNN | 0.9997415798087691 | 0.0002584201912309415 | 0.9997415798087691 | 0.9997415542813086 | NA |
| DT | 1.0 | 0.0 | 1.0 | 1.0 | NA |
| SVM | 0.9996841530996067 | 0.000315846900393373 | 1.0 | 1.0 | NA |
| LR | 0.9998277198725127 | 0.00017228012748729435 | 1.0 | 1.0 | 0.01 |

The K Nearest Neighbor model produced the best accuracy with K set to 6.

The Decision Tree model produced the best results with max depth of 3.

# Discussion

All of the models give excellent results using the test data set. The Decision Tree would be the model of choice if one model had to be chosen.

There were some surprises in the analysis of the data. *LOCATION* was not a major factor in where collisions took place. Another surprise is that *ROADCOND* was also not an influence in the severity of a collision.

# Conclusion

it is very possible to deliver a service that meets the problem statement. The data that would be immediately available from a collision could be used to warn motorists of potentially unsafe driving conditions.