Predicting Accident Probability

TAble of  
Contents

Abstract 1

Introduction 1

Problem 2

Interest 2

Methodology 3

Data Sources 3

Data Cleaning 3

Results 6

Discussion 6

Conclusion 7

# 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, ROAD\_COND, 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, Partley 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.

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.

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.

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 : Heatmap of Features

The correlation results were then sorted for SEVERITYCODE\_CAT target. Correlation results greater than 0.3 were considered relavant.

Table : 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 |
| SDOT\_COLCODE | 0.311587 |

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
* SDOT\_COLCODE

The target for the model is SEVERITYCODE.

# Results

# Discussion

# Conclusion