Heather Spero
October 1, 2020

Predicting the Severity of Accidents in Seattle

1. Introduction

1.1 Background/Problem

Every time you get in your car you have the risk of getting into an accident. In 2017, there were 187 fatal or serious accidents in the Seattle area. Obviously there are many factors that impact whether an accident occurs, including weather, road conditions, vehicle speed, time of day etc. In addition, some accidents are minor while others are very serious even fatal. The small accidents do not always block the roads, however fatal accidents can cause streets to be shut down for an extended period of time. Obviously this could significantly affect someone who had to get to a meeting or event. The problem to solve is: determine based on existing features a calculation that would predict the risk of an accident and in particular how severe it will be. Not only because of the dangers of an accident (especially severe), but also the impact an accident would have on the arrival time, etc.

1.2 Interest

This study will help potential drivers in the Seattle area determine the risk of a severe or fatal accident each time they go out, depending upon the weather conditions, road conditions and other variables which are to be determined. (The target variable is accident severity). In turn, if the accident severity shows a high probability, then these people could choose to either stay home, proceed with caution, or proceed with their drive presumably with more time. In addition, police and ambulance drivers could find this analysis very useful as they could increase the numbers of emergency vehicles on the street during times that show an increase risk of severe accidents. Finally, the city of Seattle could use the data to post warnings of dangerous driving conditions during periods of time when there is a high potential of severe accidents, in order to encourage everyone to either stay home or drive with more caution.

2. Data acquisition and cleaning

2.1 Data acquisition

The Data for this project was provided by the course and is in a csv file. It has accident data for the Seattle area from 2004 to 2020, provided by SDP and recorded by Traffic Records. The target variable that we will be analyzing will be called SEVERITYCODE and it will predict the probability of a severe accident depending on certain variables. There are 38 variable codes and approximately 200,000 entries.

2.2 Data Cleaning

There was a considerable amount of data cleaning required before the analysis could be conducted. There are 194,673 accidents reported in the file. Some of the attributes which are considered for projecting accident severity, have some missing data, including ROADCOND and LIGHTCOND, which both are missing approximately 2.5% of their values. The missing values for Road Condition and Light Condition were removed from the analysis.

The date parameters INCDATE does not have any missing data, however the date had to be translated into year (‘year’) and month (‘month’) parameters to look for seasonality or time series trends.

Most of the variables were categorical and non-numeric and therefore needed to be translated into a numeric category before any analysis could be done. One category represents road conditions at the time of the accident could be very helpful in the prediction. Therefore it had to be translated to numeric values. The following table outlines the translation categories of the variable ROADCOND to numeric values:

Dry	1
Sand/Mud/Dirt	2
Standing Water	3
Unknown/Other	3
Wet	4
Snow/Slush	6
Ice	6
Oil	6

Another category describes the light conditions at the time of the accident. The follow table outlines the translation of the variable LIGHTCOND to numeric values:

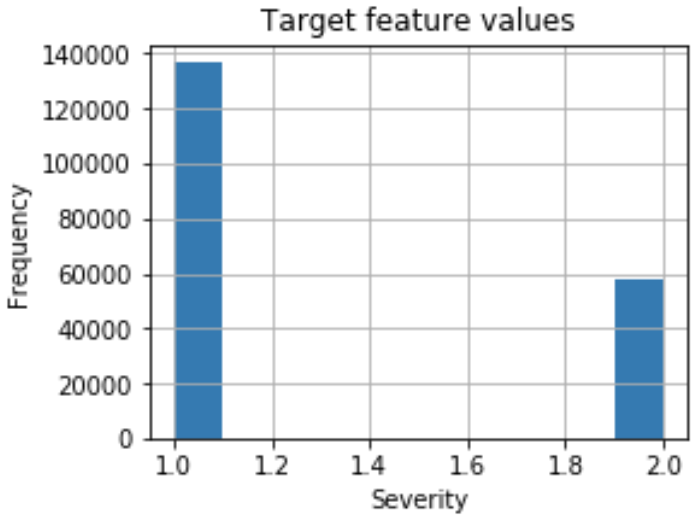
Daylight	1
Dusk	2
Dawn	3
Dark – Street Light On	3
Dark – No Street Lights	3
Dark – Street Lights Off	3
Dark – Unknown Lighting	3
Unknown/Other	4

After the data cleansing there are approximately 190,000 records for the analysis.

The four variables that are now numeric and represent the best opportunity to develop a predicative estimate of accidents include “year”, “month”, “RC_Num” and “LC_Num”. In addition, the data was standardized prior to analysis to ensure that larger or smaller code values would not distort the results.

The target value is “SEVERITYCODE” and the distribution is as follows:

The value 1 represents an accident with property damage and the value 2 represents an accident with injury. The distribution of the SEVERITYCODE is as follows:



3. Exploratory Data Analysis

Each of the variables that will be used in the analysis was analysed to determine the overall distribution of the variables.

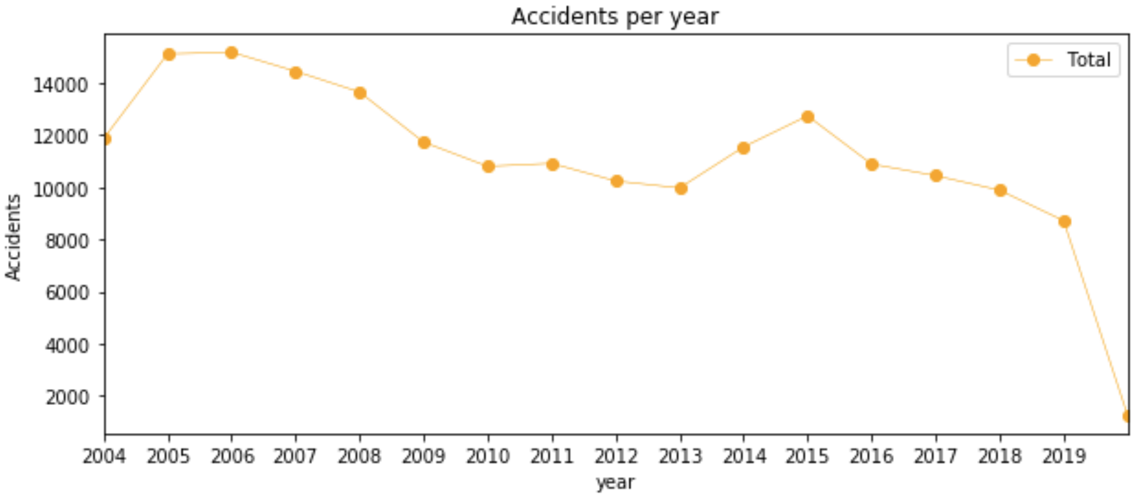
The ROADCOND (RC_Num) distribution is outlined below which highlights that the vast majority of the accidents occur in the dry weather (66%).   The large number in one category may mute the impact of this variable.

Dry	124510
Wet	47474
Unknown	15078
Ice	1209
Snow/Slush	1004
Other	132
Standing Water	115
Sand/Mud/Dirt	75
Oil	64

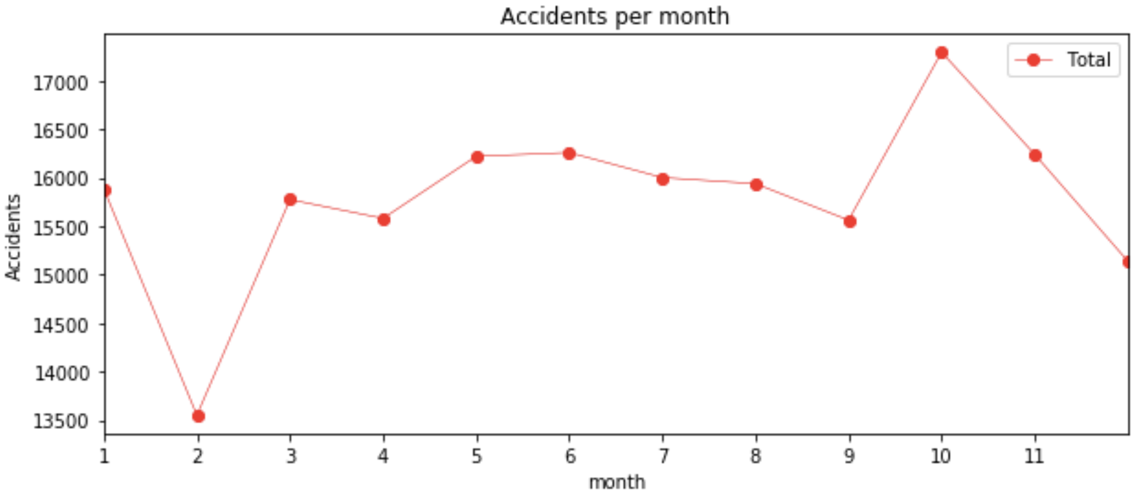
The LIGHTCOND (LC_NUM) distribution is outlined below. Similar to above, the vast majority of all accidents occur in the Daylight and therefore this variable may be limited in the sensitivity it has on the determining the Target variable.

Daylight	116137
Dark - Street Lights On	48507
Unknown	13473
Dusk	5902
Dawn	2502
Dark - No Street Lights	1537
Dark - Street Lights Off	1199
Other	235
Dark - Unknown Lighting	11

The distribution of the “year” accident data illustrates that the number of accidents per year appears to trend slightly down over the period (excluding 2020 which is an incomplete year and therefore should not be recognized in the trend). Although there is no significant fluctuation which appears as significant outliers.



The “month” data identifies 2 months that appear to be high or lower than the average. In particular October appears to be a month with an increase in accidents and February there appears to be a drop in accidents.



The correlation of each of the variables in relation to the target variable is included below:

	RD_CAT	LC_CAT	year	month
SEVERITYCODE	-.035	-.104	.023	.004

Based on the table above, each individual variable does not have a high correlation with the severitycode target value. The most impactful variable (on its own) is light conditions with a very slight negative impact, however the value is still very low and not meaningfull.

4. Methodology/Results

The cleaned data will not be put through a clustering analysis and Logistic Regression with the variables to see how the combination of the variables can significantly improve the predictive affects regarding severe accidents. The analysis programs that will be utilized are KMeans, Decision Tree and Logistic Regression.

KNN - Jaccard : 0.6858691253321134

```
          precision    recall  f1-score   support

       1       0.70      0.96      0.81     39799

       2       0.35      0.05      0.09     17034

   micro avg       0.69      0.69      0.69     56833

   macro avg       0.52      0.51      0.45     56833

weighted avg       0.60      0.69      0.60     56833
```

Decision Tree

```
Jaccard : 0.7002797670367568
          precision    recall  f1-score   support


       1       0.70      1.00      0.82     39799

       2       0.00      0.00      0.00     17034


   micro avg       0.70      0.70      0.70     56833

   macro avg       0.35      0.50      0.41     56833
weighted avg       0.49      0.70      0.58     56833
```

Logistic Regression

```
Jaccard : 0.7002797670367568

          precision    recall  f1-score   support



       1       0.70      1.00      0.82     39799


       2       0.00      0.00      0.00     17034



   micro avg       0.70      0.70      0.70     56833


   macro avg       0.35      0.50      0.41     56833


weighted avg       0.49      0.70      0.58     56833
```

The R2 score is 0 and the logloss is .6

Summary of Results Severe Accidents

Algorithm	Jaccard	F1 Score	Precision	Recall
KNN	.68	0.35	0.05	0.09
Decision Tree	.70	0	0	0
Logistic Regression	.70	0	0	0

In this case, *Precision* indicates the % of predicted accidents that were truly severe. The *recall* is the % of truly severe accidents that were properly predicted. Therefore, the recall is the most important statistic to highlight whether the predictive variables can be used to predict severe accidents accurately. In this case, in Decision Tree and Logistic Regression, the recall does not provide any support for the prediction of a severe accident. The KNN does provide a low predictive score for forecasting severe accidents. The reason the Jaccard score is so high on the predictions is because the model appears to do much better predicting small accidents.

5. Discussion/Conclusion

Based on the analysis above, it would be difficult to create a predictive model for severe accidents in Seattle based on the information provided and the variables used. Most of the data was for non-severe accidents so a small data set could be part of the problem. If someone wanted to forecast a model to predict whether small accident could occur, then the data appears to provide solid forecasting ability for that. Using the data to estimate a small accident could help solve part of the problem for some people, people who want avoid traffic jams etc., however, may not be as useful for ambulance/hospitals and police organizations to staff for serious injuries.

Algorithm	Jaccard	F1 Score	Precision	Recall
KNN	.68	0.81	0.7	0.96
Decision Tree	.70	.82	.71	.90
Logistic Regression	.70	.82	.7	1.0