Heather Spero
October 1, 2020

Predicting the Severity of Accidents in Seattle

1. Introduction

1.1 Background/Problem

In 2017, there were 187 fatal or serious accidents in the Seattle area. Globally, traffic accidents are one of the leading causes of death among young people. Obviously there are many factors that contribute to whether an accident occurs, including weather, road conditions, vehicle speed, time of year etc. It would be extremely useful to be able to utilize current information to determine whether or not an accident ill occur but more importantly whether a severe accident will occur. Many small accidents are inconsequential and do not impact the drivers or community in any significant way. However severe accidents not only impact the driver(s) involved, through injury or death, these accidents significantly impact society through blocked roads (streets to be shut down for an extended period of time), requirement of emergency services, in addition to hospital utilization.

1.2 Problem

The problem would be to try and predict when a severe accident will likely occur based on historical information. Then use the current information to forecast the likelihood of a severe accident. This could help prepare our emergency service personnel for busier times during certain conditions.

1.3 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 (1 is property damage and 2 is injury). It will predict the probability of a severe accident depending on certain variables. There are 38 variables and approximately 200,000 entries in the file.

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 that are considered for the analysis, 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 parameter INCDATE did 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:

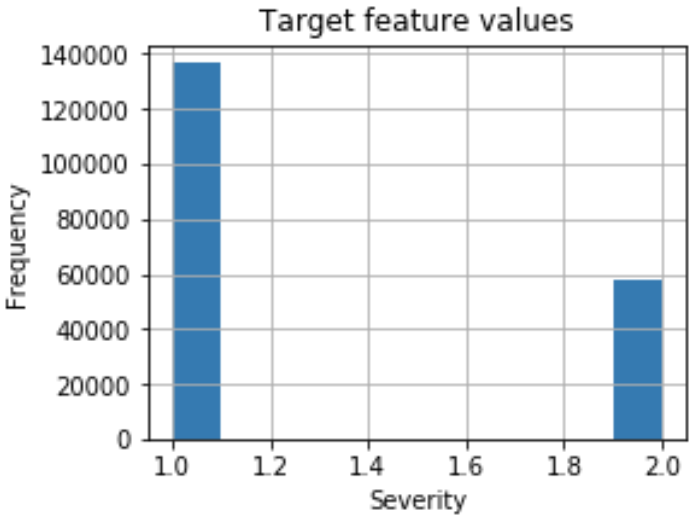
Daylight	1
Dusk	2
Dawn	2
Dark – Street Light On	3
Dark – No Street Lights	3
Dark – Street Lights Off	3
Dark – Unknown Lighting	3
Unknown/Other	2

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:



Due to the significant imbalance of the number of Code 1's versus Code 2's the dataset had to be rebalanced. This was done by downsampling the majority class (SEVERITYCODE 1's) which meant randomly removing observations from the majority class to prevent its signal from dominating the learning algorithm..

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.

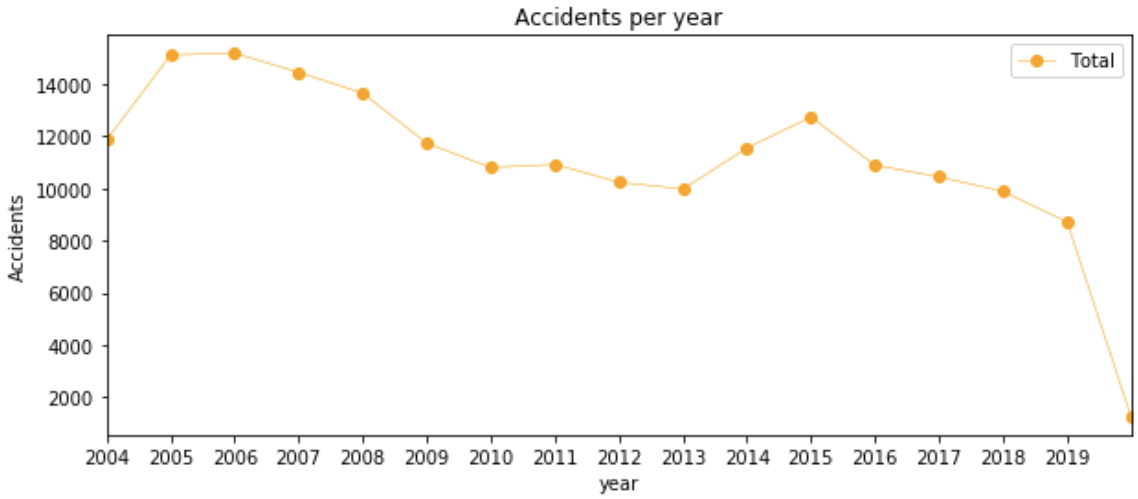
The ROADCOND (RC_Num) distribution is outlined below which highlights that the vast majority of the accidents occur in the dry weather (66%). The large number in one category may mute the impact of this variable.

Dry	124,510
Sand/Mud/Dirt	47,474
Standing Water	15,078
Unknown/Other	1,209
Wet	1,004
Snow/Slush	132
Ice	115
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.

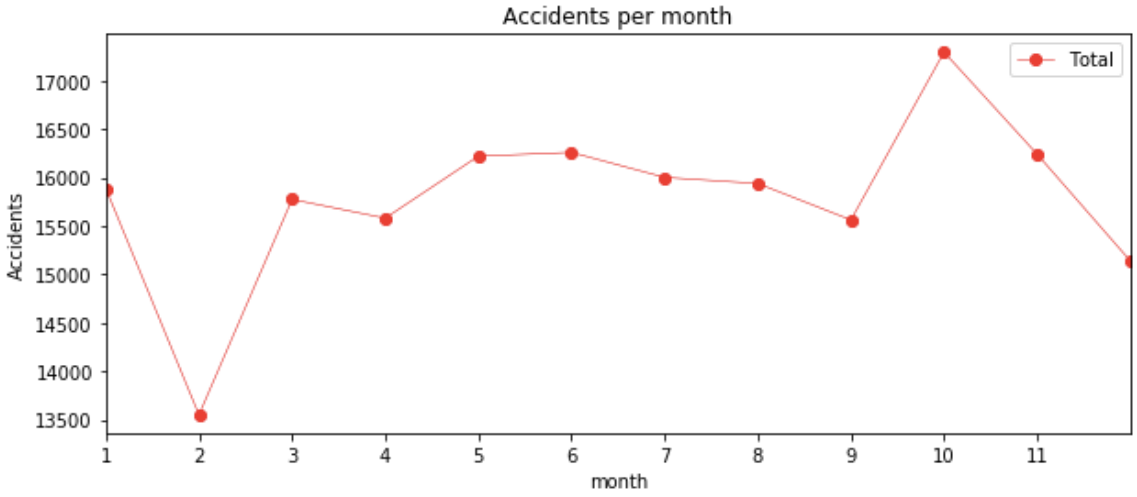
Daylight	116,137
Dark - Street Lights On	48,507
Unknown	13,473
Dusk	5,902
Dawn	2,502
Dark - No Street Lights	1,537
Dark - Street Lights Off	1,199
Other	235
Dark - Unknown	11

The yearly data was graphed in order to ensure that there is not a huge drop or outlier, especially with the older data that should be removed(drop in 2020 is because it is a partial year). 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 appears to be no significant fluctuation for any one year.



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

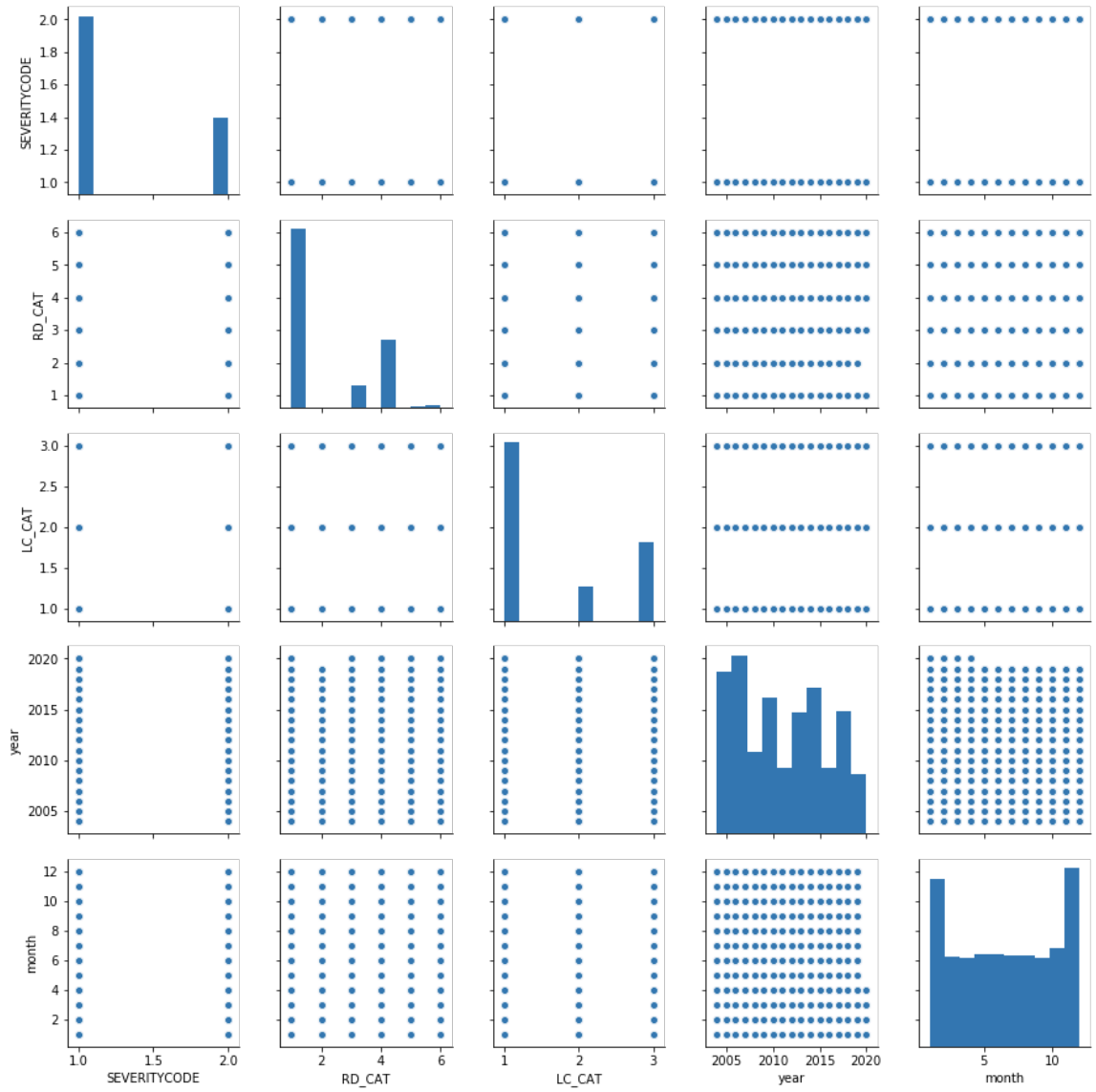
appears to be a drop in accidents. Therefore, month will be included in the forecasting analysis to help incorporate this fluctuation in the predictor.



The correlation of each of the individual 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 close to zero.



4. Methodology/Results

Different classification algorithms were utilized for the prediction for the level of accident severity(target value). The algorithms provided a supervised learning approach for predicting the amount of accuracy of the target variable. The three variables that were used in the analysis were Road Conditions, Light Conditions and Month. In order to conduct the analysis, the data was divided into two sets, 70% of the data was used for training the dataset and 30% of the data was used to test the trained algorithm. Before the analysis, the data was standardized giving zero mean and variance to all features. The different classification and logarithmic approaches are outlined below:(note: The Random Forest Model is a subset of the Decision Tree Model):

1) K-Nearest Neighbor (KNN) KNN will help us predict the SEVERITYCODE of an outcome by finding the most similar data point within k distance.

2) Decision Tree A decision tree model provides a layout of all possible outcomes so we can fully analyze the consequences of each decision. The decision tree observes all possible outcomes of different weather conditions. The Random Forest model is also used to try and improve the accuracy.

3) Logistic Regression Because our dataset only provides us with two SEVERITYCODE outcomes, the model will only predict one of those two classes(binary)therefore - logistic regression.

Summary of Results SEVERITYCODE 2

Algorithm	Jaccard	F1 Score	Precision	Recall
KNN	.53	0.50	0.53	0.47
Decision Tree	.56	0.63	0.54	0.76
Random Forest	.56	0.63	0.54	0.76
Logistic Regression	.53	0.56	0.53	0.59

Because we are concerned with predicting the serious injuries only, we will look at the F scores for value of 2 - SEVERITYCODE and the Recall Score for the value of 2 - SEVERITYCODE. 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. The Jaccard score is as important as it provides a combined accuracy for 1 and 2 SEVERITYCODES. Therefore a strong positive prediction of the 1 code would influence the Jaccard Score higher when in fact we only care about the prediction of the code 2 SEVERITYCODE.

The Random Forest /Decision Tree (Recall .76 and F1 score .63) shows significantly better results than both KNN and Logistic Regression which had very low Recall scores (.47 and .59 respectively) and much lower F1 scores as well(.50 and .56 respectively)

Therefore the best model for the purpose of predicting the severe car accidents in the Seattle area is the Decision Tree (and Random Forest) . It showed fairly impressive forecasting of the Severe accidents with a Recall of 76%.

5. Discussion/Conclusion

Based on the analysis above, it would be beneficial to create a predictive model for severe accidents in Seattle based on the historical Road conditions, Light Conditions and Month and utilizing the Decision Tree algorithm. . This model would have multiple applications in real life, including forecasting severe accidents for emergency workers and hospitals, regarding staffing etc. By utilizing current weather patterns and time of year, people could be made aware of the higher risk times of severe accidents which could help alleviate some of the problems.