IBM CAPSTONE PROJECT CAR ACCIDENT SEVERITY REPORT

Introduction

The costs of fatalities and injuries due to traffic accidents have a great impact on the society. In recent years, researchers have paid increasing attention to determining factors that significantly affect severity of driver injuries caused by traffic accidents.

Annual Global Road Crash Statistics:

- Approximately 1.35 million people die in road crashes annually, on average 3,700 people lose their lives every day on the roads.
- An additional 20-50 million suffer non-fatal injuries, often resulting in long-term disabilities.
- More than half of all road traffic deaths occur among vulnerable road users—pedestrians, cyclists, and motorcyclists.
- Road traffic injuries are the leading cause of death among young people aged 5-29.
- Young adults aged 15-44 account for more than half of all road deaths.
- On average, road crashes cost countries 3% of their gross domestic product.

Business Understanding

For problem understanding let's see the situation: You are driving to another city for work or to visit some friends. It is rainy and windy, and on the way, you come across a terrible traffic jam on the other side of the highway. Long lines of cars barely moving. As you keep driving, police car start appearing from afar shutting down the highway. Oh, it is an accident and there's a helicopter transporting the ones involved in the crash to the nearest hospital. They must be in critical condition for all of this to be happening. 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, so that you would drive more carefully or even change your travel if you are able to.

The case study is to predict the severity of an accident.

To reduce the frequency of car collisions we have to develope a model to predict the severity of an accident given the current weather, road and visibility conditions. When conditions are bad, this model will say to be more careful.

The machine learning model should be able to predict accident "severity".

Data Understanding

The dataset contains 194673 observations (rows) and 37 attributes (columns). The machine learning model should be able to predict accident "severity". The target of prediction is 'SEVERITYCODE' (it is used to measure the severity of an accident). In the dataset there are only 2 variants (1 - prop damage and 2 - injury). But the target or label columns should be accident "severity" in terms of human fatality, traffic delay, property damage, or any other type of accident bad impact. The attributes we can use to predict the severity of an accident are 'WEATHER', 'ROADCOND' and 'LIGHTCOND'.

Data Preparation:

This dataset is not fit for analysis perfectly. We should not use all attributes for our model.

Most of the attributes are text-type, so we should convert them to a numerical type. We should use label encoding to covert the features.

The target variable SEVERITYCODE is only 42% balanced. the quantity of severitycode in class 1 is 136485 and the class 2 is 58188.

There are a lot of empty fields - fill them by zero.

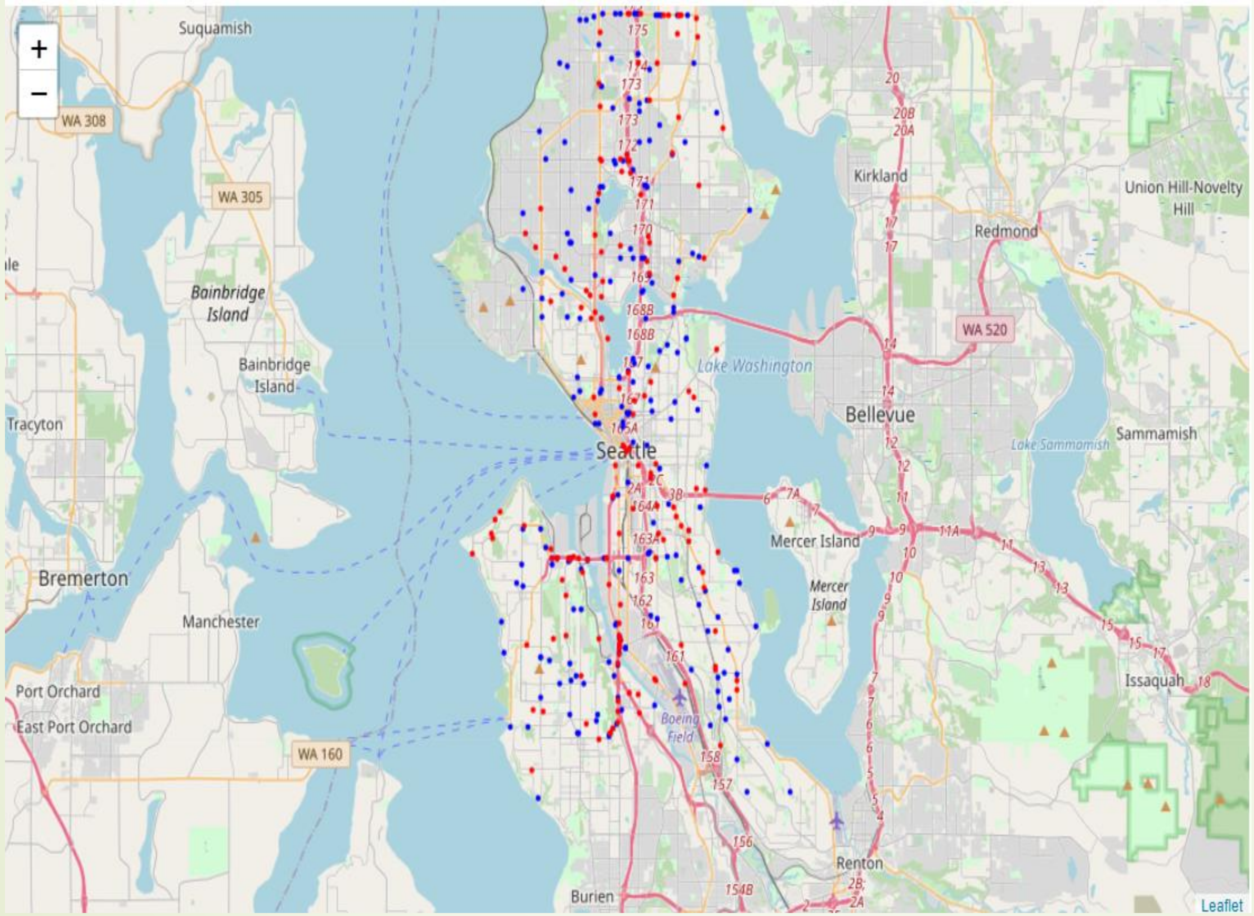
Data Analysis

Plotting factors on the map to get density of areas where accidents were caused by the features in question:

1. Speeding:
2. Under Influence (DUI)
3. Inattention
4. Hitting a parked car

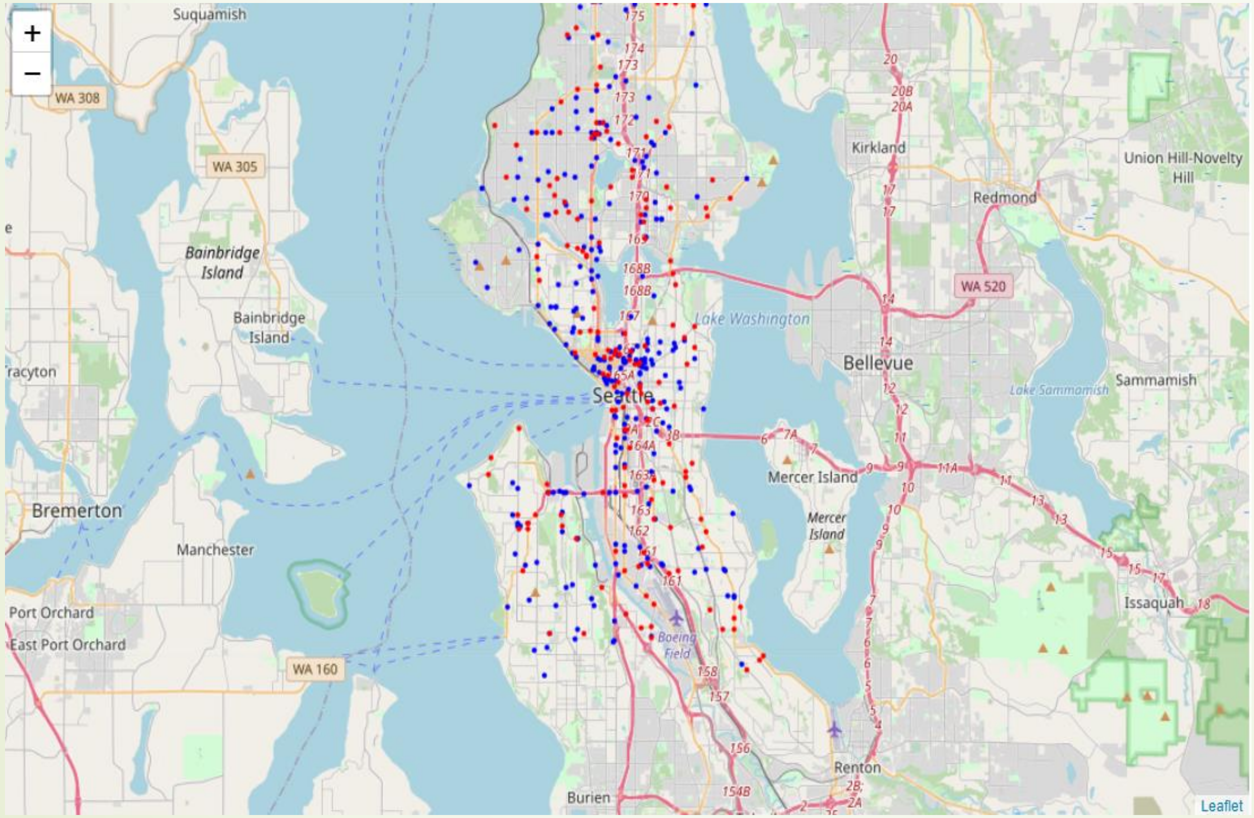
Plotting density of accidents sorted by Severity caused by Speed

Certain roads have a lot of accidents which occur on them due to speeding. The government of Seattle can introduce proper traffic management I n the form of speed restricting interventions (e.g. speed bumps). This can cause reduction in accidents due to speeding.



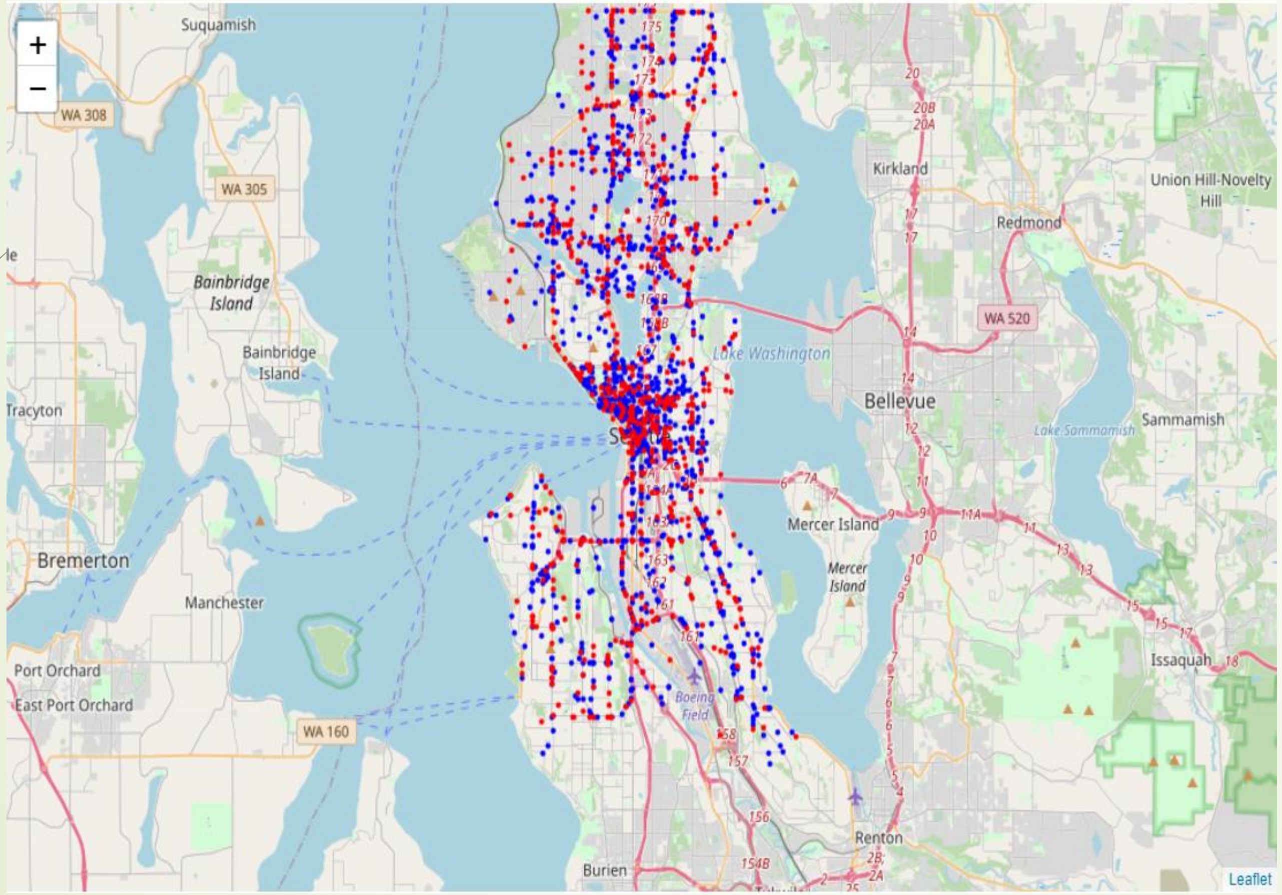
Plotting density of accidents sorted by Severity caused by Driving Under Influence

The above map shows, the points where accidents are caused due to DUI. The Seattle government can introduce Police check-ups on vehicles which are entering nodes where one has high density of accidents caused by DUI. This can reduce potential accidents before they happen.



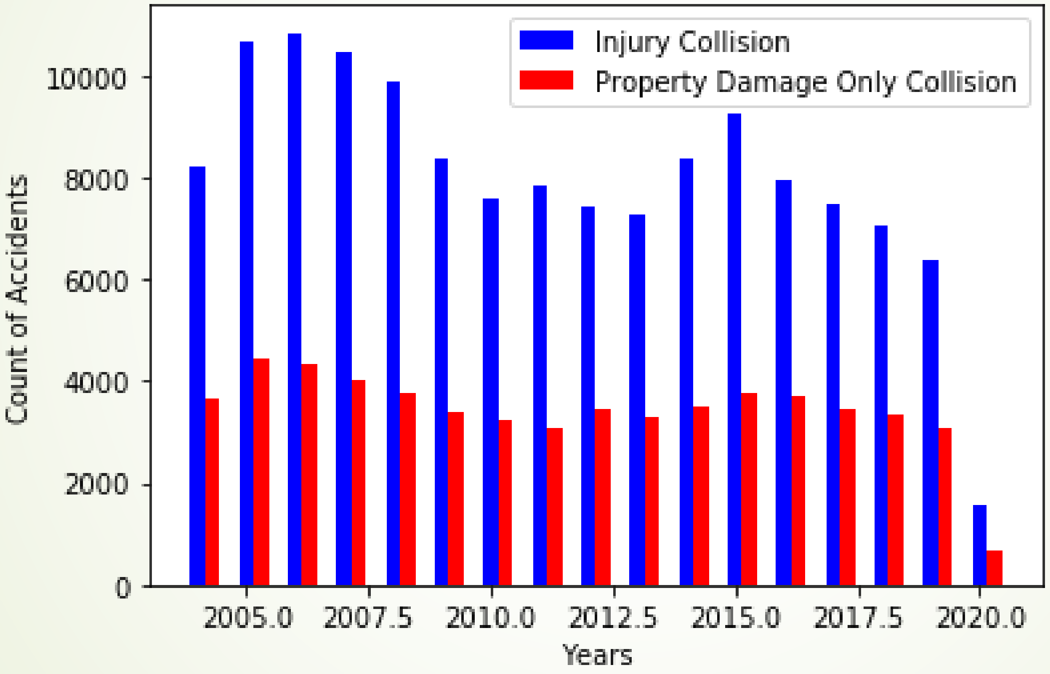
Plotting density of accidents sorted by Severity caused by Inattention

It shows that a huge majority of accidents are caused by inattention. Perhaps a product monitoring the attention of drivers can be developed/marketed citing this data.



Plotting count of accidents based on the following factors:

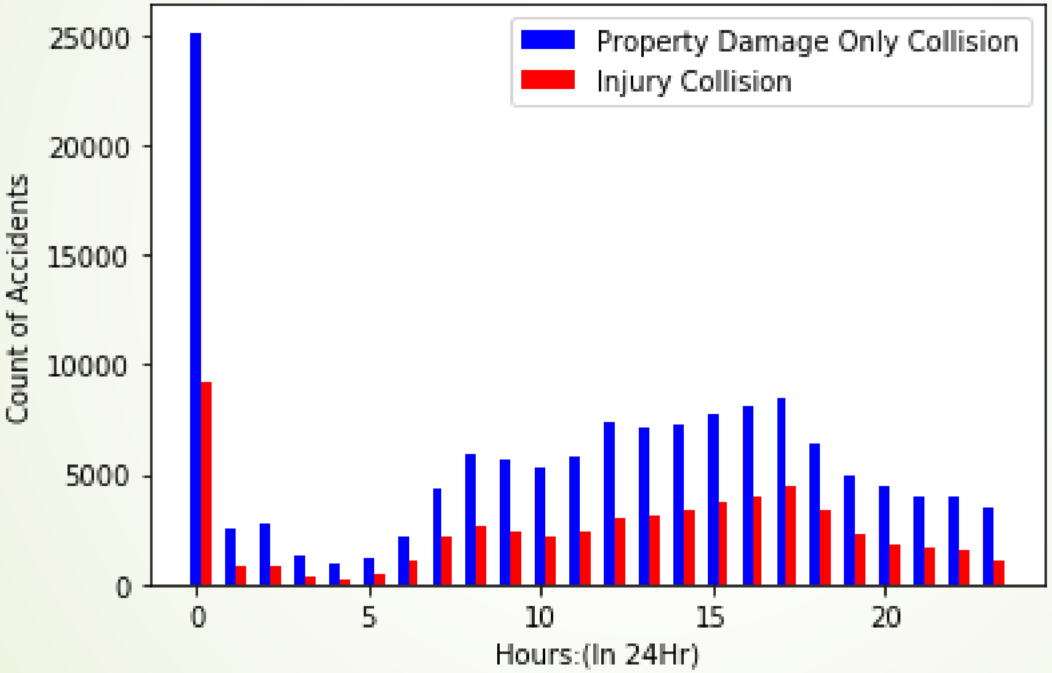
Year:



1. The Number of accidents in both the Severity classes have been decreasing over the years.
2. The drastic drop in 2020 is due to there being data for part of the year.

Plotting count of accidents based on the following factors:

Hours in 24-hour format



The highest number of accidents by far happen at midnight from 12AM to 1 AM.

Methodology

Our data is now ready to be fed into machine learning models.

We will use the following models:

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

Decision Tree A decision tree model gives us a layout of all possible outcomes so we can fully analyze the concequences of a decision. It context, the decision tree observes all possible outcomes of different weather conditions.

Logistic Regression Because our dataset only provides us with two severity code outcomes, our model will only predict one of those two classes. This makes our data binary, which is perfect to use with logistic regression.

Also we will use GradientBoosting, XGBClassifier, RandomForest and Support Vector Machine.

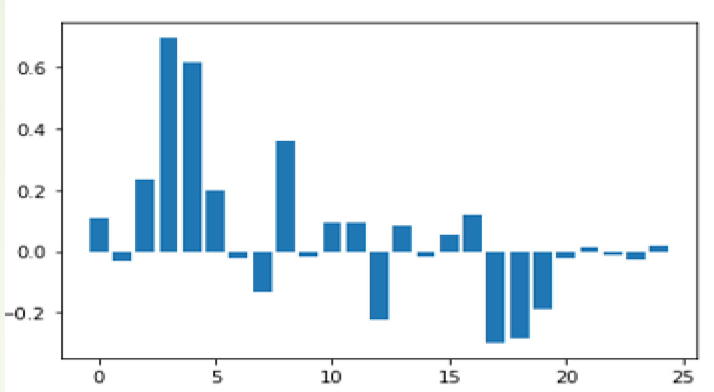
Modeling

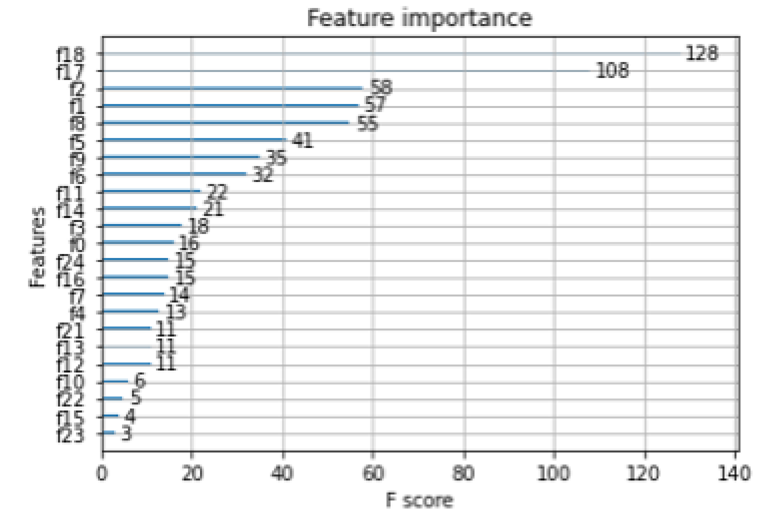
Classification models (Decision tree, K neighbors and logarithmic regression) are supervised learning algorithms that seek to classify data into a set of discrete value classes. This type of algorithm learns from the relationship between a set of categorical variables and a variable of interest, in which the latter is a categorical variable with discrete values.

Logistic regression aims to classify data from a set considering the input variables in order to predict a target variable which must be binary. It is important to note that the independent variables must be continuous.

The K-Nearest Neighbors algorithm is a classification method which groups diverse points with common characteristics to learn from the relationships between them to label unknown information. Those data that are close to each other are called neighbors.

Feature Importance Analysis





Results

Interestingly, the data shows that most accidents occur during the day with normal drivers and conditions. Most of the cases involved property damage. As expected, most occur at a block or intersection. Most vehicle accidents occur during the best driving times when it is clear, during the day, and the roads are dry.

The Results of the classification are as follows:

Model	Accuracy
K Nearest Neighbours	0.7453191216129447
Logistic Regression Accuracy	0.756876846025427
DecisionTrees	0.6912032875304995
GradientBoosting	0.7662771285475793
XGBoost	0.7648131501219982
Random Forest's Accuracy	0.7630152818800565
Support Vector Machine	0.7629639142160011

Very good results are from **GradientBoosting (best result),** XGBoost, Random Forest and Support Vector Machine.

Discussion

The first striking observation has to do with the dependent variable. It seems highly unlikely that over the date range of the data no serious injuries or fatalities occurred. This may be a warning sign that the severity codes were somehow altered when the data set was being created, or that the sample data is incomplete and missing those reports. The main recommendation has to do with key variables such as pedestrian right of way, inattentive drivers, and if the car was speeding. In many of the records, these values were null. However, this data should be collected in order to draw new insights or create better prediction models. It was determined that most accidents occur during normal weather and road conditions. However, further data is needed to analyze this trend. It may be that these types of days constitute the highest number of days in the year. Therefore, further data on the weather needs to be analyzed. For example, it may be sunny for 100 days and then snow on 1 day. If you look at data for accidents, there may be 1000 accidents occur during sunny days and only 20 on snowy days. Really, the average number of accidents per weather type is much higher on snowy days. Though, because there are few days like that during the year the total number of accidents appears low.

Once we analyzed and cleaned the data, it was then fed through three ML models; K-Nearest Neighbor, Decision Tree and Logistic Regression. Although the first two are ideal for this project, logistic regression made most sense because of its binary nature. Evaluation metrics used to test the accuracy of our models were jaccard index and f-1 score. Choosing different k, max depth and hyparameter C values helped to improve our accuracy to be the best possible.

Conclusion

The data-set has been used to classify the severity of the accidents based on certain select features. The exploratory data analysis shows density of accidents based on geography based on Speeding, Driving Under Influence, In-attention and Hitting Parked Cars. From a machine learning standpoint. The most important features were: Collision Type, Person Count, Vehicle Count and Address Type. The Gradient Boost algorithm performed the best.

The data showed that most vehicle accidents occur during good conditions with normal drivers. This means it will be harder for the Seattle transportation department to mitigate accidents. However, as most accidents only involve property damage or minor injuries, there is not a serious problem that needs to be dealt with right away. This shows that infrastructures are being designed and operating properly. Therefore, the focus should be an emphasis on drivers being more careful.

Based on historical data from weather conditions pointing to certain classes, we can conclude that particular weather conditions have a somewhat impact on whether or not travel could result in property damage (class 1) or injury (class 2).