Emergency Medical Services Alert System (EMSAS)
Ruben Goncalves
November 2020

Introduction / Business Problem

Accidents happen! Road accidents are a risk most people have to contend with at various points in their lives. They can be anything from a simple property damage situation to a fatality as well as various degrees of injury.

However road accidents are made worse by their unpredictability and while its important to make sure there are enough police officers on duty to attend and process every incident, it is vital that a sufficient number of paramedics and emergency support staff is present when needed.

There is thus a need for a model that based on easily observable conditions such as the weather, light and road conditions can predict if a serious incident is likely to occur to alert the Emergency Medical Services (EMS) which can prepare paramedics in advance and improve readiness.

Data

The dataset comes from the city of Seattle - available at : https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv - where the target variable SEVERITYCODE can vary between 0 and 4. However the sample only includes entries for CODE 1 – Property Damage and CODE 2 – Injury, which is fine since for the purpose of the model every code other than CODE 1 would need to alert the paramedics and as such would be included under Injuries (CODE 2).

The dataset provides many variables for analysis but not all of them are useful. Even if they could be used for general analysis, they can not be used for prediction. For example, if the data were to suggest that accidents between a pedestrian and a car lead to a higher number of injuries (CODE 2) than property damage (CODE 1) it is of limited predictive value since the number of pedestrians at each accident cannot be determined beforehand in a useful and expedient way – and so it cannot be used to alert the Emergency Medical Services (EMS).

As such, a more limited number of variables will be chosen and several machine learning models will be applied to the cleaned dataset. The one with the best accuracy will be deployed to alert the Emergency Medical Services (EMS) and help save lives.

Methodology

In order to obtain useful insights that could be used to improve the alertness level of the EMS it was necessary to clean the available dataset from any features that are not of use. An initial data exploration revealed many features that possess correlation with the target variable (SEVERITYCODE), however most of them can not inform our business interest since they are descriptive in nature, i.e. they can only be determined after the accident has taken place. An extreme example of this is the variable SDOT_COLCODE and its descriptive pair SDOT_COLDESC, which together describe an accident in detail and assign an numerical code to the accident. It is logical that a collision between a truck and a bicycle is more likely to result in a Class 2 entry (Injury) in the target variable than a collision between two bicycles and, as such, these features are correlated with the target variable. Unfortunately, since one can not know in advance which type of

accident will occur, it has no predictive value and must be dropped. Many other variables are also by nature descriptive and must be dropped. The following table lists all dropped variables and the reason for being removed from the dataset.

Table 1. - Variables removed from the data set and the reason why.

Variable removed from the dataset	Reason for the removal from the dataset
OBJECTID, INCKEY, COLDETKEY, REPORTNO and SDOTCOLNUM	Are for internal use of the city of Seattle and are unique identifiers that convey no information.
EXCEPTRSNCODE, EXCEPTRSNDESC	Have no metadata to describe the features and STATUS is similarly undocumented.
SEVERITYCODE.1 and SEVERITYDESC	Are duplicates of the target variable and convey the same information.
PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT	While may be useful for analysis post incident, they are descriptive in nature and can not be used for prediction as one can not in advance know many people, pedestrians, bicycles or vehicles will be involved in an accident.
COLLISIONTYPE, JUNCTIONTYPE, INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SPEEDING, SEGLANEKEY, HITPARKEDCAR	Have the same limitations as the features above since one can not know if people will be inattentive, under the influence, if they grant a pedestrian right of way, are speeding, on which lane they are, the type of junction or collision when accident happens or if they hit a parked car until after the accident.
SDOT_COLCODE, SDOT_COLDESC, ST_COLCODE and ST_COLDESC	Are highly descriptive features that while possessing plenty of useful data, they can not be used as a basis for an alert of the Emergency Medical Services.
CROSSWALKKEY, INTKEY and ADDRTYPE	Duplicates - Information is already present in the coordinate variables X and Y.

The variables INCDATE and INCDTTM were then broken down into useable variables. INCDATE was broken down into YEAR, MONTH and DAY and INCDTTM was broken down into HOUR.

The resulting dataset was then limited to the variables SEVERITYCODE, X,Y , YEAR, MONTH DAY, HOUR, WEATHER, ROADCOND and LIGHTCOND.

The variables SEVERITYCODE, WEATHER, ROADCOND and LIGHTCOND were then transformed from categorical features into numerical by the use of indicator variables.

To avoid an unbalance of classes in the models used, the classes were balanced by randomly removing an excess of class 1 SEVERITYCODE entries (downsampling).

Temporal visualisation was performed for the total cases and the class 2 entries against the feature HOUR in order to see if the distribution was uniform and if the class 2 entries followed a different pattern than the total incidents.

Geographic visualisation of class 2 incidents was used through the Folium library against the Seattle Map and a further visualization was done regarding the distribution of this entries by time, during the night period (between midnight and 5 am) and the day period (between 6 am and 11 pm).

After cleaning and downsampled the dataset was also normalized through the standard scalar method and a train test split into 2 groups was performed at the ratio of 4/5 for the Training set and 1/5 for the Testing set.

The dataset was then submitted to four models: KNN (K-Nearest Neighbours), Decision Trees, Logistic Regression and SVM (Support Vector Machines) and all relevant features were used. This includes all the Road, Weather and Light Condition features as well as the TIME features excluding YEAR (since the year never repeats and cannot be used for prediction).

Finally a simple analysis of frequency of the accidents was performed, specifically the proportion of class 2 incidents of the total incidents was determined, how many injuries happen per day based on the YEAR 2019 (most recent and relevant complete year) and how this translates to the expected number of injuries at any one time during the day.

Results

1) Relation by YEAR

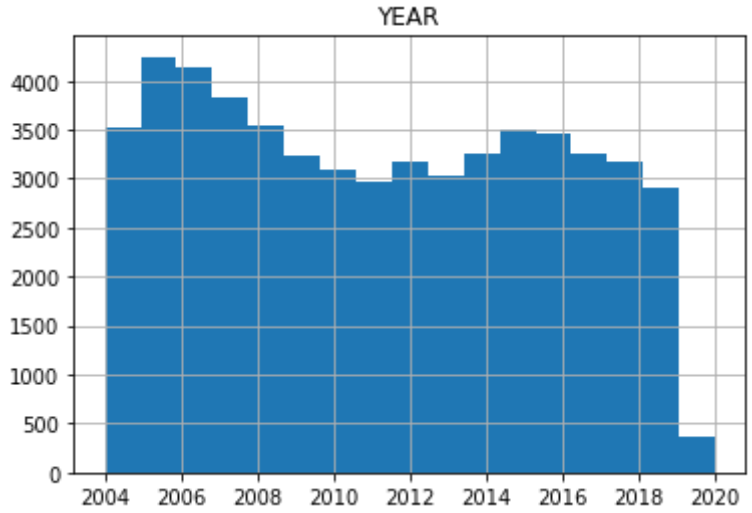

Figure 1. - Analysis of dataset of Total number of Injuries (Class 2) based on the Year the accidents occurred.

It is important to note that while the year 2020 is represented in the dataset as determined per the value_counts() method it is not complete since 2020 has not finished and care not be used interpreting that result. Also there appears to have been a slight downward trend that has stabilized of late.

2) Relation by MONTH

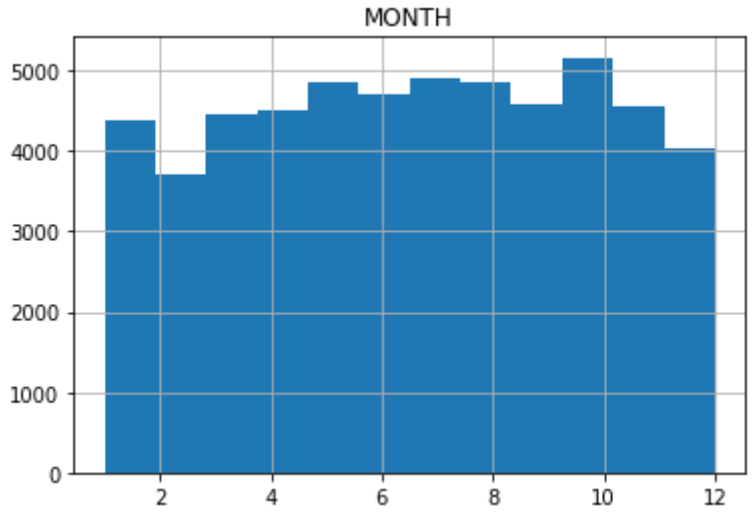

Figure 2. - Analysis of dataset of Total number of Injuries (Class 2) based on the Month the accidents occurred.

In here it becomes apparent there is no significant variation per month of the number of injuries.

3) Relation by DAY

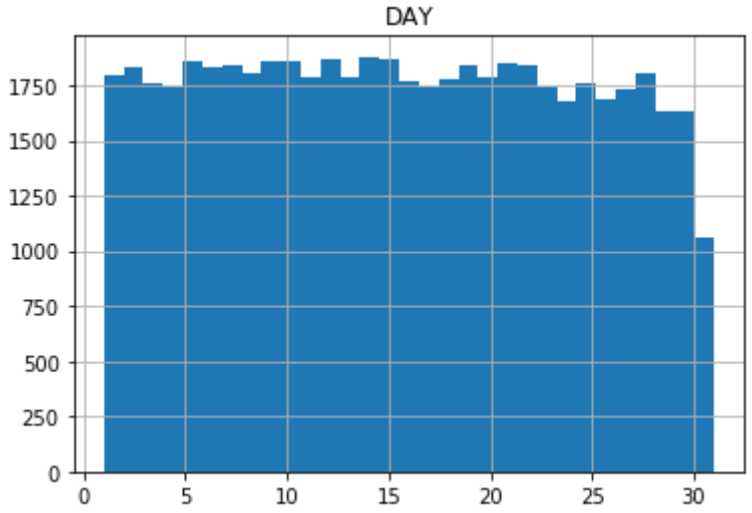

Figure 3. - Analysis of dataset of Total number of Injuries (Class 2) based on the Day the accidents occurred.

The 31st day has half the injuries since only about half of the months of the year have 31 days. Apart from this there is no significant variation per day of the number of injuries.

4) Relation by HOUR

4a) Total number of entries

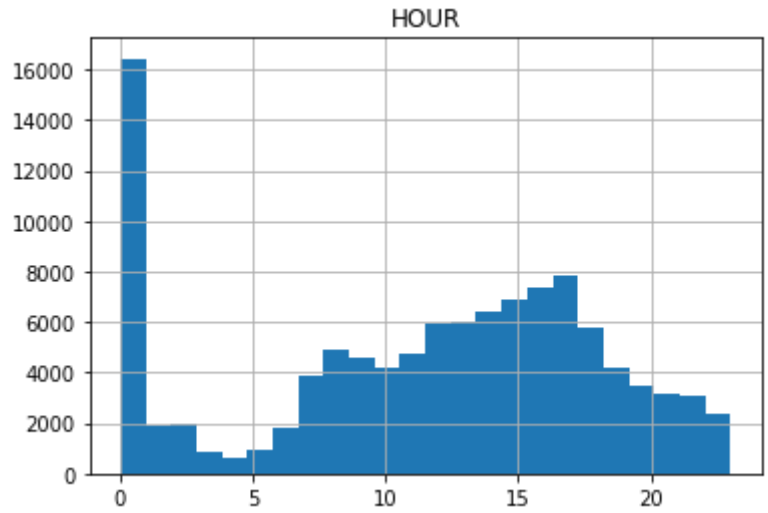

Figure 4. - Analysis of dataset of Total number of Accidents (Class 1 +Class 2) based on the Hour the accidents occurred.

4b) Class 2 entries

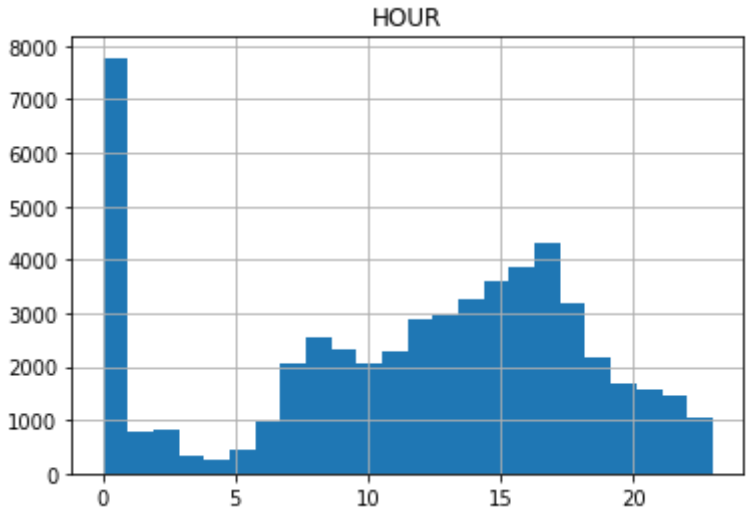

Figure 5. - Analysis of dataset of Total number of Injuries (Class 2) based on the Hour the accidents occurred.

It becomes apparent there is a significant variation regarding the time the accidents happen being a spike near midnight and the majority of cases occurring during the daylight hours. However there does not seem to be any significant difference between when the injuries (Class 2 entries) occur and the overall accidents (Class 1 + Class 2).

5) Geographic Visualization

5a) Injuries (Class 2 entries) geographic distribution

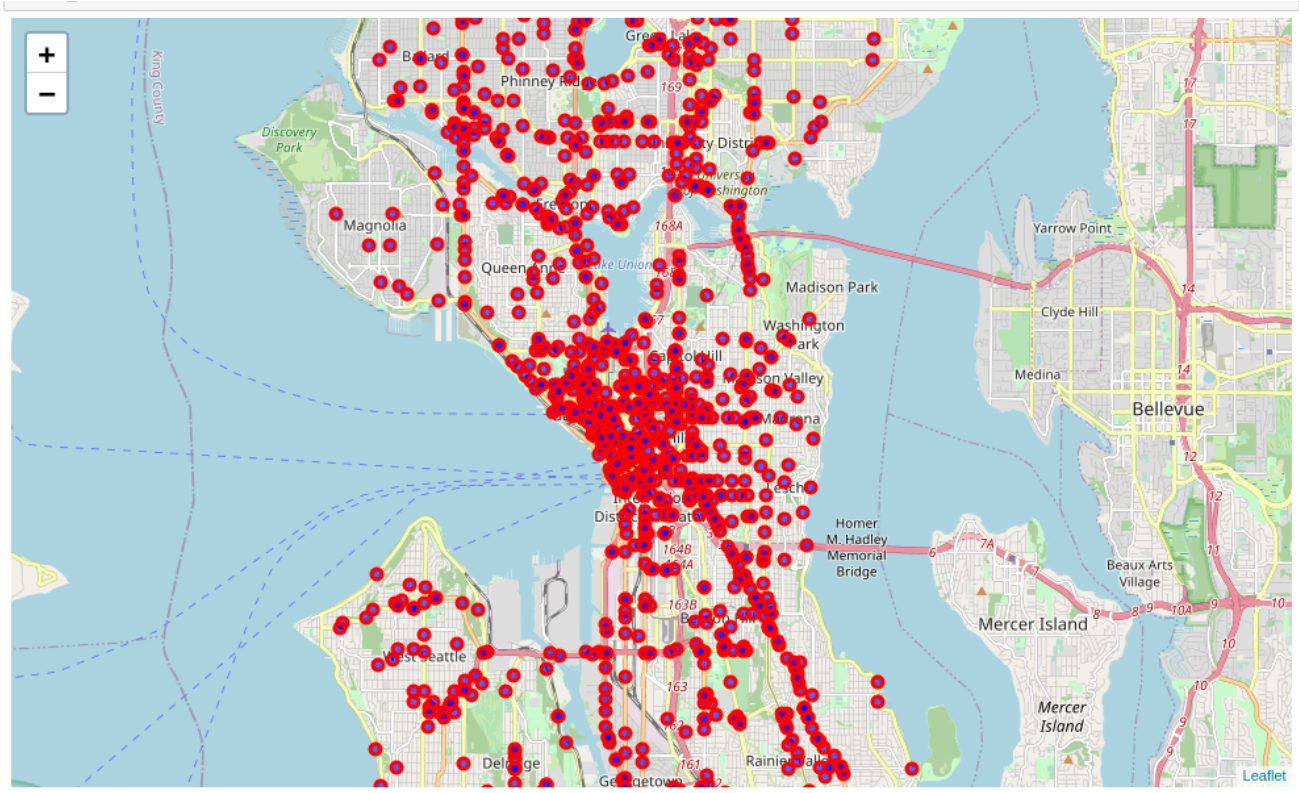

Figure 6. - Visual representation of the injuries (Class 2 entries) on a Seattle map. Number of entries limited to 1500.

It is apparent from the geographic distribution that most injuries occur clustered at the city centre.

5b) Injuries (Class 2 entries) geographic distribution by time of day

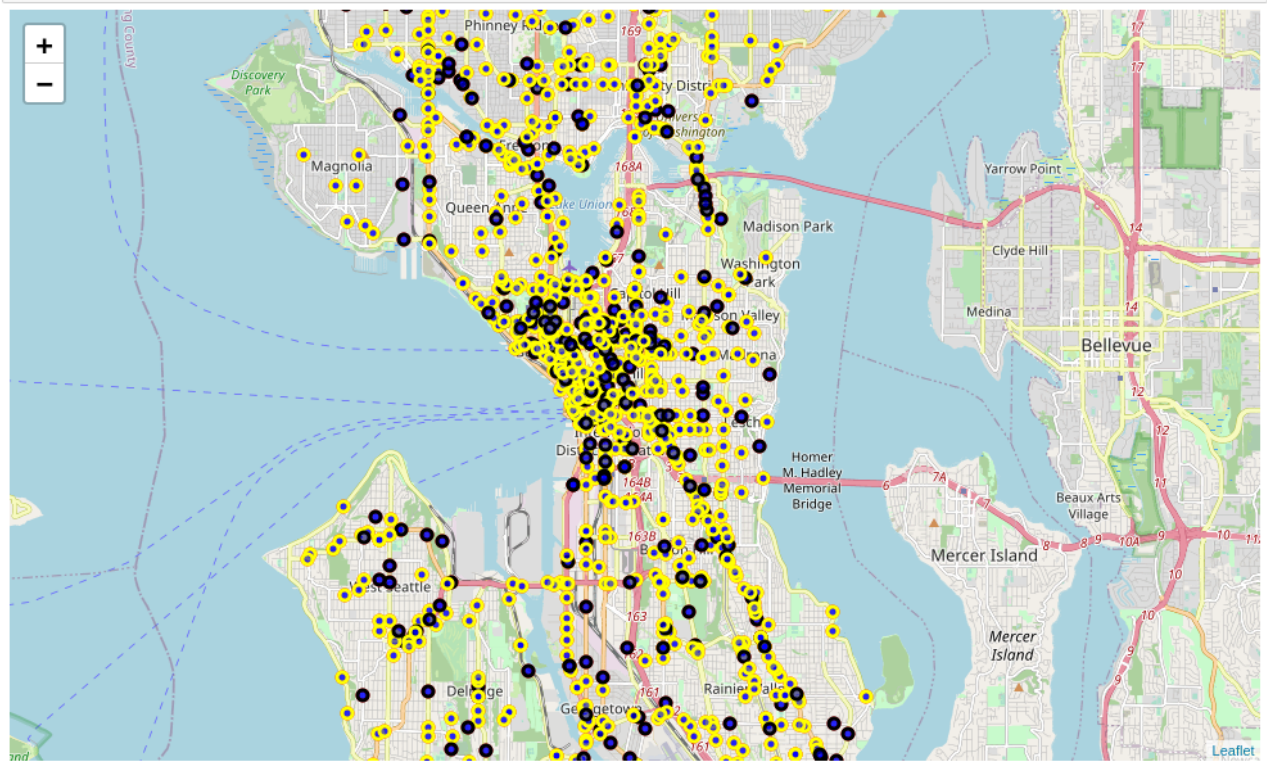

Figure 7. - Visual representation of the injuries (Class 2 entries) on a Seattle map by period of the day. Number of entries limited to 1500. Blue dots – accidents that happen between midnight and 5 am. Yellow dots – accidents that happen between 6 am and 11 pm.

It is apparent that while most injuries happen during the day there is no significant difference in the geographic distribution of the injuries by period of the day. They are both clustered around the city centre. This means it has more to do with overall traffic most likely and not a specific district more frequent during the night for example for drinking purposes.

6) Model Accuracy

6a) KNN (K-Nearest Neighbours)

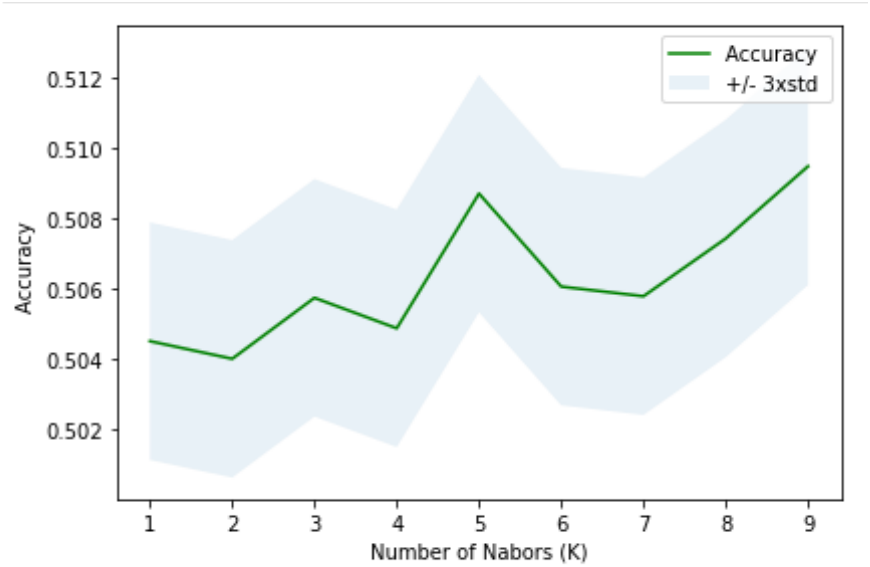

Figure 8. - Graphic representation of the KNN model accuracy for K between 1 and 9.

Table 2. - Classification report for the KNN model. K = 9.

	precision	recall	f1-score	support
1	0.51	0.50	0.51	10895
2	0.51	0.51	0.51	10990
micro avg	0.51	0.51	0.51	21885
macro avg	0.51	0.51	0.51	21885
weighted avg	0.51	0.51	0.51	21885

It is apparent that the precision of this model at identifying Class 2 entries is inadequate since even at K = 9 it is only 0.51. Considering a random coin toss would return a precision of 0.50 it means it is only about 2% better than random guessing.

6b) Decision Tree

Table 3. - Classification report for the Decision Tree model. Max depth =10.

	precision	recall	f1-score	support
1	0.51	0.57	0.54	10895
2	0.52	0.46	0.49	10990
micro avg	0.52	0.52	0.52	21885
macro avg	0.52	0.52	0.52	21885
weighted avg	0.52	0.52	0.52	21885

It is also apparent that the precision of this model at identifying Class 2 entries is also inadequate since even at MaxDepth of 10 it is only 0.52. Considering a random coin toss would return a precision of 0.50 it means it is only about 4% better than random guessing.

6c) Logistic Regression

Table 4. - Classification report for the Logistic Regression model.

	precision	recall	f1-score	support
1	0.52	0.43	0.47	10895
2	0.52	0.60	0.56	10990
micro avg	0.52	0.52	0.52	21885
macro avg	0.52	0.52	0.51	21885
weighted avg	0.52	0.52	0.51	21885

The Logistic regression model precision suffers from the same issue as the other models although its recall is slightly better. It is unfortunately still inadequate for deployment.

6d) SVM (Support Vector Machine)

Table 5. - Classification report for the SVM model.

	precision	recall	f1-score	support
1	0.52	0.35	0.42	10895
2	0.52	0.68	0.59	10990
micro avg	0.52	0.52	0.52	21885
macro avg	0.52	0.52	0.50	21885
weighted avg	0.52	0.52	0.50	21885

Finally, the SVM model also proves inadequate with a precision of 0.52 however it has the best recall of the tested models at 0.68.

7) Frequency Analysis

Incidence was determined using the formula:

Incidence = (number of class 2 entries / (total number of entries))
= (58188/(136485+58188))
= 0.29890123437764865

Number of daily accidents was determined using the data from 2019 as it was most recent complete year available. The formula used was:

Accperday = (8246/365)
= 22.59178082191781

Number of injuries per day was determined using the formula:

Injperday = Accperday*Incidence
= 6.752711174460523

Table 6. - Expected number of injuries per hour at any given day.

	TOTALINJURIESPERHOURPERDAY
0	0.130206
1	0.105639
2	0.110553
3	0.077796
4	0.059780
5	0.085985
6	0.137576
7	0.328382
8	0.404540
9	0.369327
10	0.340666
11	0.380792
12	0.432383
13	0.416824
14	0.484793
15	0.438116
16	0.444667
17	0.513455
18	0.381611
19	0.285799
20	0.242397
21	0.222743
22	0.205546
23	0.153136

Discussion

The nature of the business needs was to create a model to predict accidents with injuries when certain road and weather and road conditions happened. Easily observable features such as month, day and hour could also be used to predict when the EMS would be needed. After cleaning the data and analysing it using the four models, it became apparent that there was no significant advantage in using the variables with information about the weather, light, time and road conditions for classification purposes. The models showed that it is not possible to accurately predict accidents with injuries based on these conditions alone since they have a precision of around 0.52 (as shown in Tables 2 through 5) which is barely better than a coin toss. Prediction would thus have to be dependent on previous observations and would not be prediction at all but descriptions of an accident. Thus, it seems that there is no viable way to predict when accidents with injuries will happen based on the particular conditions of that day.

However, the data showed a relation between the location and the frequency of an accident as well as between the hour and the frequency of accidents. A simple frequency analysis was then applied. The proportion of class 2 over the total entries (1 and 2) is 29.8 %. The average number of daily accident is 22.6 and the average number of injuries per day is 6.75. A hourly distribution of the injuries was also calculated (Table 6). Geographically it was also clear that the majority of accidents happen in Seattle's city centre (Figure 6). Thus the emergency services can expect the number of accidents present in Table 6 at each hour mostly clustered at the city centre and plan accordingly.

Conclusion

An estimate of the number of injuries per hour was found as well as that accidents are clustered near the city centre which can be used to plan and inform alertness levels. No significant relation was found between road condition, light conditions and weather in determining if an accident will result in injury or not.