**Introduction**

**Background**

Seattle Department of Transportation (SDOT), Traffic Management Division collects the data on the car accidents and has the information since 2004. The data includes all types of collisions and is updated weekly. It consist of the type of accidents, whose parties were involved, the severity of collision, location, weather, date, time and etc.. There are almost 200,000 entries.

**Problem Description**

The severity of the accidents can depend on several factors, such as the condition of the roads, visibility on the roads depending on the day time and weather, the vehicle speed, and whether the driver was under effect of alcohol and /or drugs, any other distractions. Some of these decisive factors are determined by the driver and some are not.

**Interest**

One of the approaches for the SDOT to prevent the car accidents on the roads is to clearly define factors that affect the severity of the car accidents on the roads, which are not in the scope of influence of the driver. The SDOT can minimize collisions by informing the drivers ahead of the potential risks on the roads.

**Goals of the project**

The purpose of this project is to predict the severity of the accident based on the:

- External Factors

- Human Factors, such as speeding, not paying attention to the road and under influence of drugs and/or alcohol

- Physical aspects of collision sight, such as address type, junction type and road conditions

- Time Series Analysis

We can predict danger levels of this influencing factors, while being openminded to any new questions that will come up during the analyses.

**Data Acquisition & Cleaning**

**Data Sources**

All the data is provided by SDOT Traffic Management Division, Traffic Records Group and starts at 2004 till present day.

**Data Cleaning & Feature Selection**

There are lots of missing data from the Human factors. The Speeding section seems like a very important factor. However, it has only about 5% data values, so we will drop it.

Inattention column there is only about 15% values of the whole dataset.

Also in Under Influence category there are different values Yes/No and 0/1 that I have changed into only Y/N. The missing data will need to be considered later on in the analyses.

For the external factors, one of the features we will use is Lighting condition. About 7% of the accidents don’t have information for the light conditions that I will combine with ‘other’ category in the light condition. There are 2 types of dark light that needs to be combined. While for the weather conditions there are about 8% of the unknown data that I will combine with ‘other’ category.

Address type, Junction type and Road conditions all have categorical data and are characteristics of the collision site. Road conditions column is missing about 7% of the data.

Some of the above data are categorical and some are numerical, we will need to adjust it along the way to perform the analyses.

The Weather feature will also be selected since it can impact the drivers visions and change the surface of the road.

We will also us Severity level of the collision. There are 4 types of severity, however our database shows only 2 types. Where severity type 1 indicates property damage and type 2 indicates injury.

We replace all the not available data with “N” to be able to evaluate the data set.

In our further analyses, we will use time series. So for convenience we will create a columns with the year, month, day of the week and hour of the collision. It will help us to see if there are any peaks in collisions during particular time and day.

There are is complete data for the year 2020, so we will drop the whole year 2020.

Some of the above data has numerical values and some have blank spaces which will need to be addressed as well.

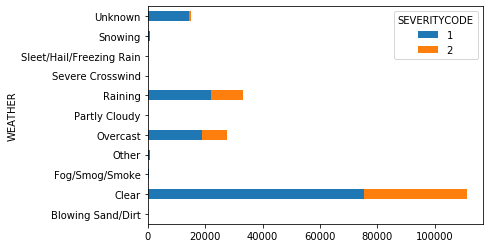
The rest of the data we will drop for now to simplify the work.

**Exploratory Data Analysis**

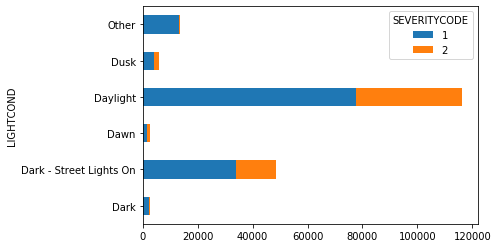
In this section I will show the relationship between several features of the available data set and accidents, particularly the severity of the collision. This will help us to better understand which predictive factors will need to be used in the Modelling section of this report.

**External factors**

Looking at number of accidents in relations to weather conditions, we can see that most of the accidents happened on the clear day for both severity types. Does it mean that more traffic is happening on a clear day? This feature cannot be the main predictor for our analysis.

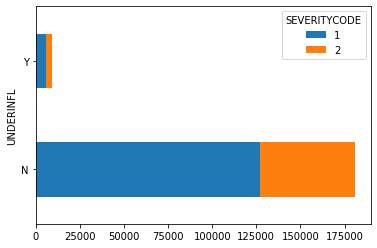
****

The graph below show the distribute of collisions based the lighting conditions. Again most of the cases happened during the day light and the second category in relations to the number of cases happened in the dark but with the street lights. Considering that there is lots more traffic during the day, we can say that lighting conditions can direct us to better estimate car accidents.

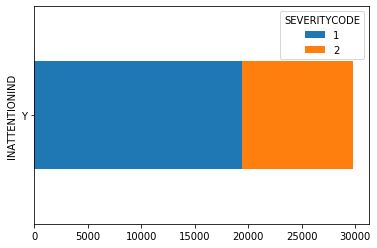
****

**Human Factors**

The below chart show two columns. The column “Y” shows number of cases where the driver was under the influence of alcohol and /or drugs. The second column “N” show were the driver was sober. Again it is an important factor that impact the accidents but in this case it does not clear indicate that the influence of alcohol or drugs is the main predictor of collision.

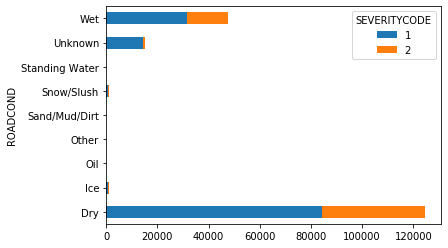
****

There are not much data available on whether the drive was paying attention to the road or not. Only about 15% of the cases got confirmation that the driver was not paying attention. If we assume that the rest of the case the driver paid attention then this factor is also not very good predictor for future collision. It has minimum impact.



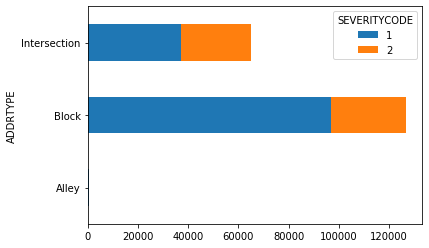
**Characteristics of Collision Site**

The conditions of the road is a very similar feature to the weather. Since the rain causes the roads to be wet, which is the second most common category in the weather graph below after dry roads. We can say it does not fully explain the number of accidents.

****

Address type data shows that most of the accidents have happened at the block, the second most common category was an intersection.

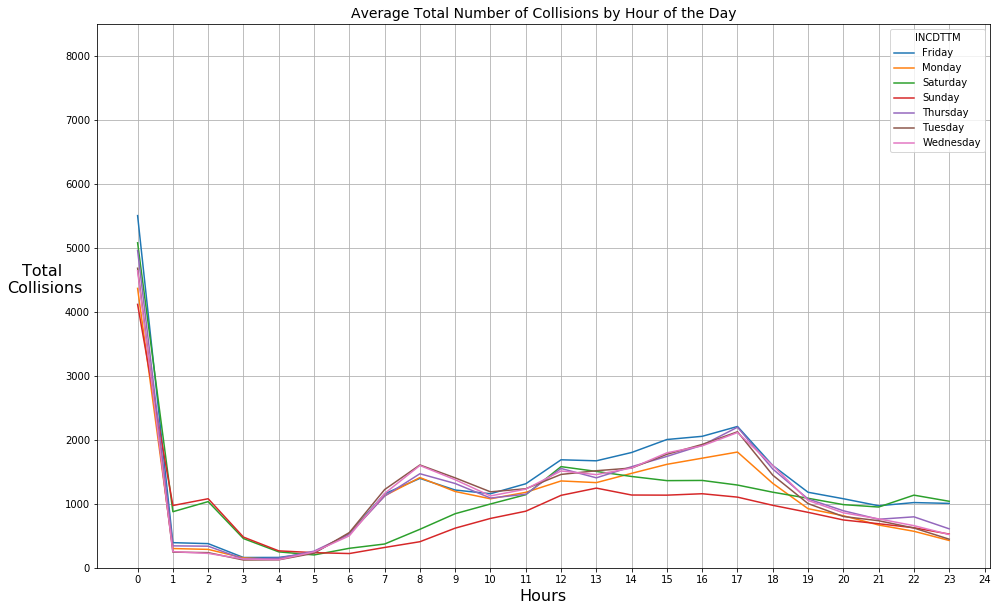
The Junction type shows similar results as the address type, just splits the categories into more detailed sub sections. So we will not use it in our analyses.

****

**Time series analysis**

I have created separate features “month”, “day of the week” and “hour” for the purpose of these analyses using the time and date of the accidents feature.

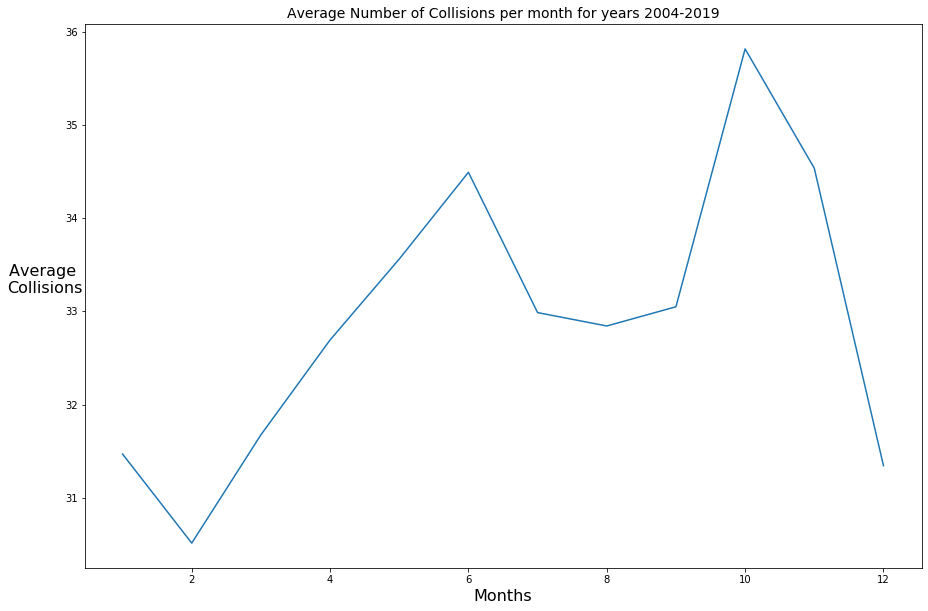
Here we are looking at the distribution of the accidents depending on the time of the day and it is very obvious that there is a big spike in the number in the middle of the night. So time of the day is a good predictive factor for further analysis.

****

The table below shoes the total number of accident in years 2004-2019 happened per day of the week. The highest number happens to be on Friday while the lower on Sunday. This is also important predictor as it shows clear dependence and we will include it in final modelling.

|  |  |
| --- | --- |
| **Friday** | 32333 |
| **Monday** | 26338 |
| **Saturday** | 27389 |
| **Sunday** | 21955 |
| **Thursday** | 29324 |
| **Tuesday** | 28556 |
| **Wednesday** | 28778 |

If we look into the average number of collisions per month of the years 2004-20019, it is noticeable that there is a big rise in number that starts in September and peaks in October, then drops end of the year. Maybe it has to do with the weather getting conditions getting worse of people resuming driving after holidays. In any case, it also an important independent variable for building our model.

****

**Predictive Modelling**

In this section we will analyse how above mentioned conditions impact the severity of the collisions. We only have enough suitable data for two categories outcome: severity 1, where the property was damaged and severity 2 where injury took place. These will be our two classes/outcomes.

We will be using various classification algorithms. All the data will be split into training and testing sets. First we will train the model and then test it, which is supervised learning techniques. One of the final part of the analysis will be obtaining the occuracy of each of the classification algorithms and then compare their accuracies to each other to find the best model.

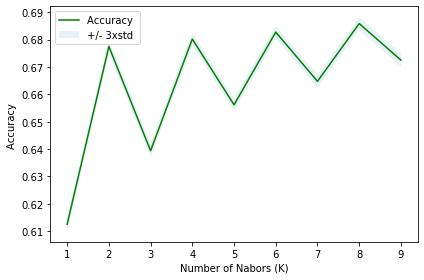
To prepare the dataset for the supervised algorithms learning, I have replaced all categorical values with numerical using LabelEncoder method. After that all the data was normalized using StandardScaler.

The final step before applying the algorithm was to split the data into training and testing set for the analyses.

**K-Nearest Neighbours with Time Series Analysis**

This method is the most obvious to start with, since it classifies the cases based on their similarity to each other.

To pick the right K number we calculate the accuracy for the K values 1-10.



In the above graph we can see that the highest or optimum accuracy is when K=8. So we will use it for our analyses.

**Other Algorithms with Time Series Analysis**

We will also turn to Decision Trees, Logistic Regression and Support Vector Machine algorithms.

The table below shows the performance of the classification models. The accuracy for all of these methods is shown and the highest ones are selected in red.

| **Algorithm** | **Jaccard** | **F1-score** | **LogLoss** |
| --- | --- | --- | --- |
| KNN | 0.67 | 0.64 | NA |
| Decision Tree | 0.7 | 0.59 | NA |
| SVM | 0.7 | 0.58 | NA |
| Logistic Regression | 0.7 | 0.59 | 0.59 |

**Modelling without Time Series Analysis**

The model with the dataset without day of the week, month and hour shows very similar accuracy for the algorithms. With the highest number being 0.7.

| **Algorithm** | **Jaccard** | **F1-score** | **LogLoss** |
| --- | --- | --- | --- |
| KNN | 0.69 | 0.63 | NA |
| Decision Tree | 0.7 | 0.59 | NA |
| SVM | 0.7 | 0.58 | NA |
| Logistic Regression | 0.7 | 0.59 | 0.59 |

**Results**

The results can be categorised by the accuracy of the predictive models. The highest number was 0.7, which is highlighted in red in the tables above.

**Discussion**

The data available for this project is complete. Lots of crucial features are missing data, such as speeding. Also for the outcome of the collision the was only data for the severity of property damage and injury. So only 2 outcomes. It will be more interesting to create the model with higher number of outcome.

Also considering that the technology of the cars and traffic is consistently changing, it might also be important to consider doing the analyses. Only using the data from the past 5-7 years. Since previously, people didn’t use the mobile phones as much in the car, navigation systems and techniques were completely different. Plus the cars themselves were not as advance as nowadays.

**Conclusion**

We looked at the Seattle Department of Transportation dataset for the collisions for the years 2004-2019. First, the data was analysed to see which features from the dataset could be affecting the severity of the accidents. The second step involved classification method of the machine learning to predict the severity of the accident based on the human factors, external factors, description of the potential collision site and time of the day, day of the week and month. The highest accuracy for the models achieved was 0.7, which is reasonably good.