Predicting Accident Severity of Seattle City

Ashutosh Tiwari

September 07, 2020

1. **Introduction**
   1. **Background**

Seattle city is a [seaport](https://en.wikipedia.org/wiki/Port) city on the [West Coast of the United States](https://en.wikipedia.org/wiki/West_Coast_of_the_United_States). It is the [seat](https://en.wikipedia.org/wiki/County_seat) of [King County](https://en.wikipedia.org/wiki/King_County,_Washington), [Washington](https://en.wikipedia.org/wiki/Washington_(state)). Seattle is the largest city in both the [state](https://en.wikipedia.org/wiki/U.S._state) of Washington and the [Pacific Northwest](https://en.wikipedia.org/wiki/Pacific_Northwest) region of [North America](https://en.wikipedia.org/wiki/North_America). Because of being the coastal area the movement of high transport is common on daily basis and due to which the safety of people are more important in this area cause on accident can hold the large portion of city steady for a while. SO it is helpful for the city to predict the accident occurrence and its severity so that lots of life can be saved and people can be aware of the problems they are going to face ahead under certain conditions.

* 1. **Problem**

The data that might be helpful for this project has to contain the conditions under which the accident happen and what type of accident and its severity took place we need to have the location where this happened , road conditions, lightening conditions , weather etc. to help us predict the accurate severity of accident.

* 1. **Interest**

The Seattle transport department are the main department can take interest in it for better handling the traffic and accident occurrence, can help the people of Seattle city or to those who are travelling to the city by providing the safety precautions under the conditions they are driving on road or sea ships for better management.

1. **Data acquisition and cleaning**
   1. **Data sources**

The Seattle data provided by the Seattle city transport department can be found[Clicking here](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv). And meta data  [Clicking here](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf) which contains the relevant information and descriptions for norms used in dataset.

This dataset contained the data from 2004 to 2020 for the incidents of Seattle city under different coordinates.

* 1. **Data cleaning**

The dataset contains the several unnecessary data for prediction of severity goal hence while identifying those data found out that the severity code were mentioned twice so we need only one column for this project hence we removed the column. The data had null values to it for some records hence we removed the data for better scenario understanding. The column which did not had any relevance to our goal were also removed such as ["SEVERITYCODE.1","OBJECTID","INCKEY","COLDETKEY","REPORTNO","INTKEY","EXCEPTRSNCODE","EXCEPTRSNDESC","INATTENTIONIND","PEDROWNOTGRNT","SDOTCOLNUM","SEGLANEKEY","CROSSWALKKEY"]

Because these are more connected to the description of records for the incidents happened and maintaining records.

The Data we had some of them are outer then the Seattle city coordinates hence we removed those data and finally the total number of records we got are (184167, 25). The redundancy in the data were removed while handling the formatting of data before doing any analysis on it.

* 1. **Feature selection**

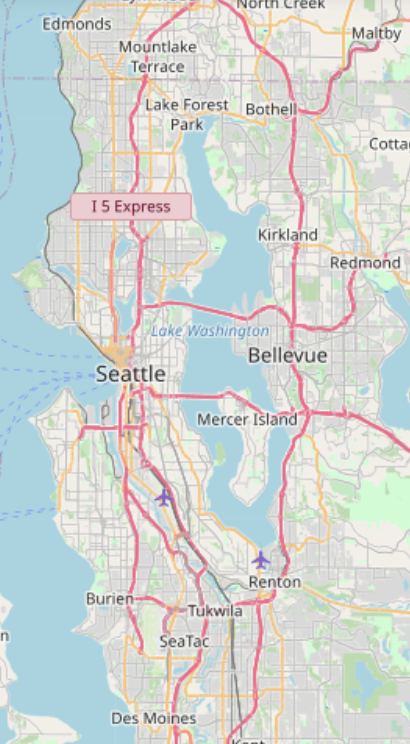
After data cleaning, there were 184167 samples and 25 features in the data. Upon examining the meaning of each feature, it was clear that there was some redundancy in the features. For example, the Junction type had the blank values in it which cannot be used under the scenario of model designing hence handled it by deleting such rows, the coordinates X,Y are under the object data type hence to convert them under float values for better implementation of map and analyse the data for it.

Table 1.1 Feature selection and dropping reason map

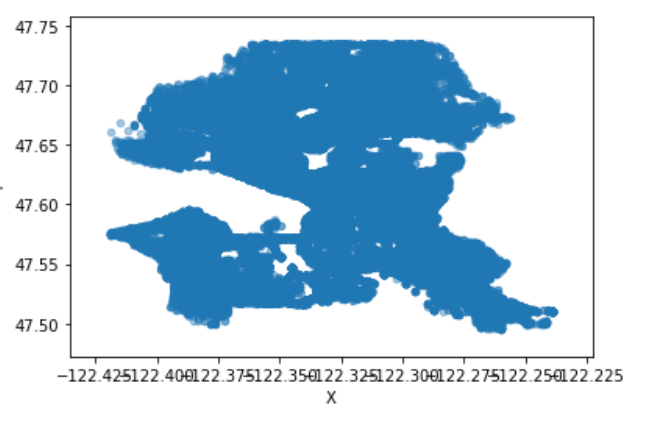
|  |  |  |
| --- | --- | --- |
| Feature selected | Feature Dropped | Reason for dropping |
| “X” | “LOCATION” | This is dropped as the coordinates X,Y are used to identify the location and this parameter was duplicating the records. |
| “WEATHER” | “INCKEY” | This was meta data for records maintenance and not needed for severity analyse |
| “ROADCOND” | “INCDATE,  INCDTTM” | This was dropped for having same date records and can be used for detailed review but not prediction |
| “LIGHTCOND” | “SDOT\_COLCODE” | Not required for this project |
| “UNDERINFL” | “SDOTCOLNUM” | Not required for this project |

1. **Exploratory Data Analysis**
   1. **Seattle city Location Vs Crowd**

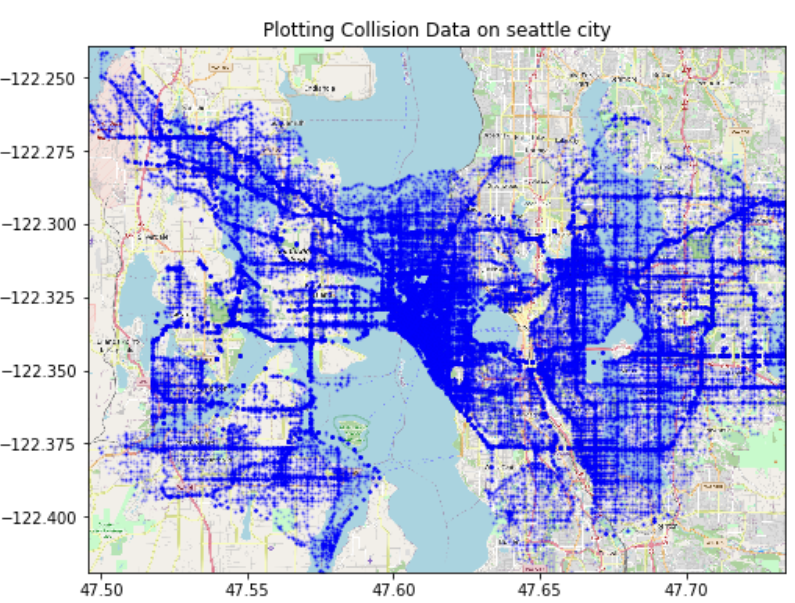
The coordinates are provided in the dataset for analysing the data of severity in Seattle city. Which can easily be plotted and can be analysed for better understanding to data on USA map for severity.



Lets analyse the data on an scatter plot which will help to find any pattern and analytics for the severity and location of accidents happened in Seattle city for the year 2004 till 2020 and how it will impact the target people for specific location data.



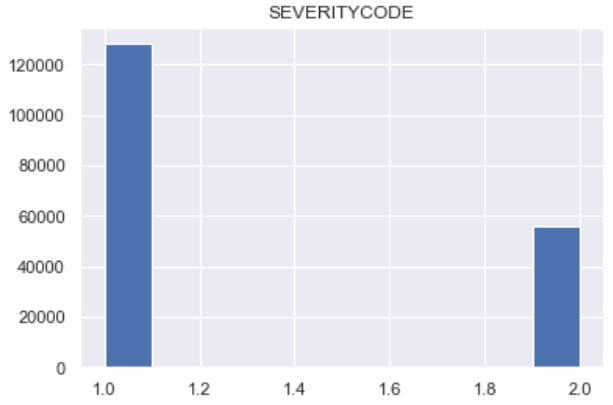
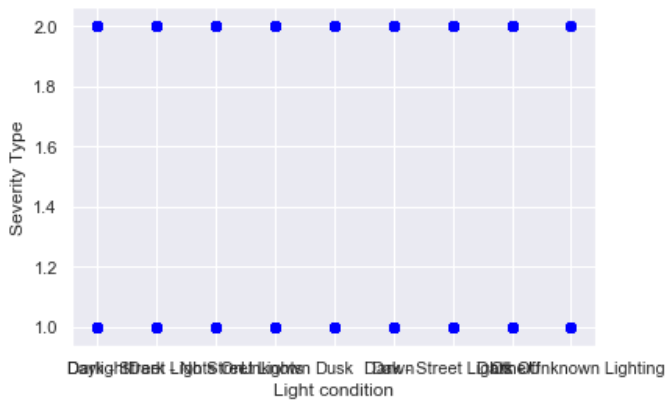
As we can clearly observe in the data plotted there are no specific patter we can observe but the cluster plotting will not clearly defines the nature of data how it can help the models to analysing the accident severity, for better understanding on the data of scatter plot we can use the map on the scatter plot to analyse the actual impact on different locations. As shown below.

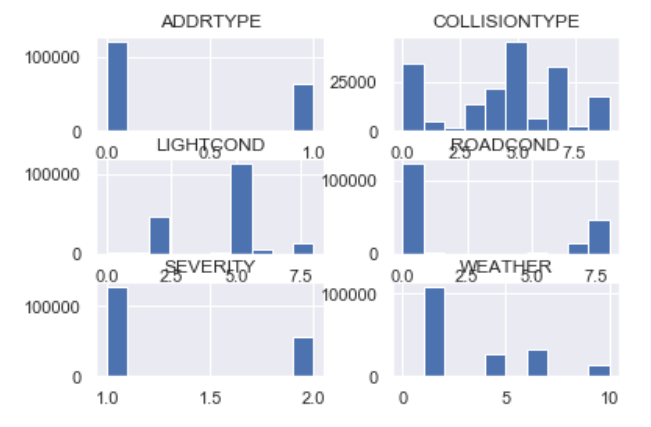


As we can clearly observe in the above plot that some areas are highly populated with the accidents and while observing closely we can see these are the coastal areas which are heavily crowded and local market contains the most of population, these needs to be monitored closely.

* 1. **Relationship between Severity and light conditions**

The analysing of the severity with the context of different feature we use the most common were is the light conditions for which the accident severity type happened the below mentioned graph will show the severity type that happened under lightning conditions as there are several different type fused for these scenarios.



The different features we used for the data analysis will help to better perform the task of moving towards the goal as we can clearly see in the above graphs the different features will impact the accident occurrence on different severity levels.

As the COLLISIONTYPE graph shows how this some type of collisions are the most common among the accident severity and this will impact he data on high basis.

The road conditions are mostly negatively impacted the accident severity as we can see is the graph of “ROADCOND” the impact is mostly 0 and if it moves forward it is 0.75 which can not impact the accident on high level.

1. **Predictive Modelling**

There are two types of models, regression and classification, that can be used to predict accident severity. Regression models can provide additional information on the amount of accident, while classification models focus on the probabilities a accident severity category or type. The underlying algorithms are similar between regression and classification models, but different audience might prefer one over the other. For example, the transport department of Seattle city might be interested in calculating the risk of severity of accident and the people of Seattle city might be interested in the possibility of accident occurrence.

**4.1 Regression models**

**4.1.1 Applying standard algorithms and their problems**

I applied linear models (linear regression, Ridge regression, and Lasso regression), Logistic Regression to the dataset, using root mean squared error (RMSE) as the tuning and evaluation metric. The results all had the same problems. The predicted values had much narrow range than the actual values, and as a result, the prediction errors were larger as the actual values deviated further from zero. These results were not acceptable, because accidents with large severity were arguably more important for Seattle city department to predict than accidents with little change in severity. Having larger errors on those predictions was obviously not desirable. 4.1.2 Solution to the problems The reason behind these problems were the uneven distribution of accident severity. My solution to this problem was to assign weights to samples based on the inverse of the abundances of target values. In other words, accidents with large severity rate would have higher weights in model training and evaluation because they were more rare. Using this method, all models predicted target values with similar range and distribution as the actual target values

**4.1.2 Solution to the problems**

The reason behind these problems were the uneven distribution of accident , in that severity were more common than accident with high severity. Therefore, the models tried to prioritize minimizing errors on accidents with little severity class when RMSE was used as the evaluation metric. My solution to this problem was to assign accidents to samples based on the inverse of the abundances of target values. In other words, accidents with high severity rate would have higher weights in model training and evaluation because they were more rare. Using this method, all models predicted target values with similar range and distribution as the actual target values.

**4.1.3 Performances of different models**

Using the new approach of different sample light and collision type, I built linear regression, logical regression models using weighted root mean squared error as the evaluation metric. For each model, hyperparameters were tuned using the same metric and cross validation. For comparison, I also built a simple linear regression model with just one independent variable (COLLISIONTYPE) as the benchmark model. K nearest model had the best performance among all models, which had k=4 less error than the benchmark model. The predicted improvements had k nearest relationship with the actual improvements .

Table 2. Performance of the regression/classification models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Simple regression** | **Decision Tree** | **K nearest model** | **Logistic Regression** |
| Accuracy | 0.42 | 0.724 | 0.726 | 0.688 |

1. **Conclusions**

In this project, I analysed the relationship between road accident severity and their geographical and conditional data. I identified road condition, lights, weather conditions ,collision type and mostly the location among the most important features that affect accident severity on road. I built both regression models and classification models to predict whether and how much accident severity impacts on it. These models can be very useful in helping Seattle city transport department management in a number of ways. For example, it could help identify accidental prone areas, weather conditions warning to the people , road condition monitoring etc.

1. **Future** **directions**

During project was able to achieve ~42% improvement from the benchmark model in the regression problem, and ~72% accuracy in the classification problem. However, there was still significant variance that could not be predicted by the models in this study. I think the models could use more improvements on capturing traffic data traits. The severity of different accidents ac also be further calculate or analysed for example suppose we have a different accidents happens at the same intersection can lead to the more prediction rate and further we can take more precautions. More data, especially data of different types, would help improve model performances significantly. Models in this study mainly focused on conditional features.