**Predicting the severity of accidents**

**1. Introduction:**

**1.1 Background:**

**Car accident is one of the most critical life changing action. It affects not only the driver but also the whole society.** **Road traffic injuries are currently estimated to be the eighth leading cause of death across all age groups globally, and are predicted to become the seventh leading cause of death by 2030. beside of that, it costs millions of dollars yearly. Hence, it is important to find a model to predict the severity of accident and the relationship with surrounding condition. So, The Government and other related service organizations can take proper actions to prevent sever accidents and prepare resources which are needed to react adequately. Machine Learning is the best choice to predict required info in real time regarding multi-factorial and sophisticated environment.**

**1.2 Problem:**

**This project aims to predict how much severity of accidents is based on information such as: address type, junction type, inattention of driver, drugs or alcohol usage, weather condition, light condition, whether pedestrian right of way, speed, existing of parked car, where collusion occurred lane segment or crosswalk or intersection.**

**1.3 Interest:**

**Government agencies, healthcare service provider, and Travelers would be very interested in accurate prediction of severity.**

**2.Data understanding:**

**2.1 Data sources:**

**To resolve this problem, we need a dataset contains information about accidents and the factors which may lead to them such as the Traffic Records Group from Seattle.**

**2.2 Data cleaning:**

**There were lots of missing values (<5% except SDOT\_COLCODE 15.3%) and I deal with it as follow:**

* **Drop severity code missing values**
* **For INATTENTIONIND,** **SPEEDING,** **PEDROWNOTGRNT feature, missing values are considered as “No” values**
* **The missing values in other features are replaced by mean or mode values according to type of variable and suitable classes regarding severity code.**

**Some values such as others or unknown are identified as missing values and replaced with best value as described previously.**

**2.3 Feature selection:**

* **Features which may help us :( Location, address type, junction type, inattention of driver, drugs or alcohol usage, weather condition, light condition, whether pedestrian right of way, speed, existing of parked car, where collusion occurred lane segment or crosswalk or intersection, type of collision; count of persons, pedestrians, vehicle, and bicycles involved)**
* **severity of these accidents: (severity code).**
* **Omitted features: ‘SDOT code’, 'OBJECTID', 'INCKEY’,'INTKEY' ,'COLDETKEY' ,'REPORTNO' ,'STATUS' , 'LOCATION' ,'EXCEPTRSNDESC' ,'EXCEPTRSNCODE' ,'SEVERITYCODE.1' ,'SEVERITYDESC' ,'INCDATE' ,'SDOT\_COLDESC' ,'SDOTCOLNUM' ,'ST\_COLDESC',** **'SEGLANEKEY','CROSSWALKKEY'.**

**Some of these features contain texts to describe the severity or location, others contain Id keys for records, or special codes such as SDOT.**

**3. Exploratory Data Analysis:**

**Here I will describe some examples how some features related to severity with charts:**

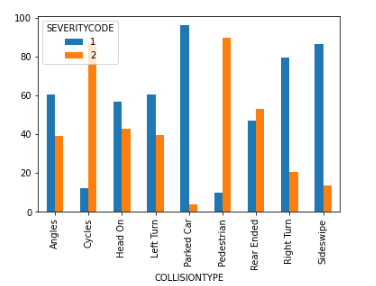
**3.1 Relationship between severity and Address type & junction type:**

**We notice that intersections and junctions related to them have more sever accidents (chi square p<5%).**

|  |  |
| --- | --- |
|  |  |

**3.2 Relationship between severity and Collision Type:**

**Angels, Cycles, Head On, left turn, Pedestrian, Rear ended have more severity (chi square p<5%).**



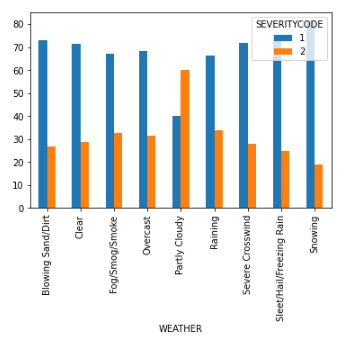
**3.3 Relationship between severity and Inattention, Drugs, and alcoholism:**

|  |  |
| --- | --- |
|  |  |

**Also, they are combined with more sever accidents (chi square p<5%)**

**3.4 Relationship between severity and Weather:**

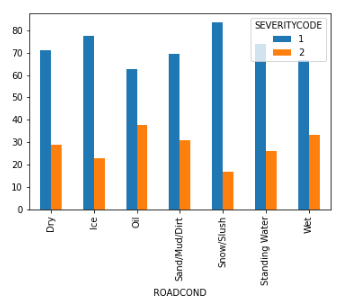
**Bad weather conditions (such as fog, wind, cloud, rain. etc.) lead to more sever accidents (chi square p<5%).**



**3.5 Relationship between severity and Road condition:**

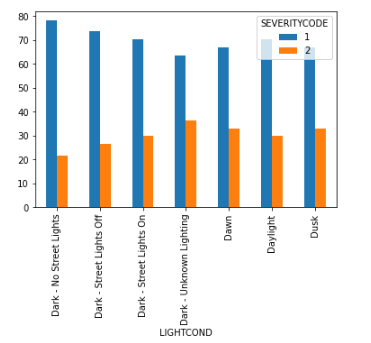
**We notice that there is an increase in severity related to wet, oil, and dry.**

**Paradoxically, snow and slush have lower severity but this is not meaning that the number of accident lower. this is about the presence of severity which is our concern here (chi square p<5%).**



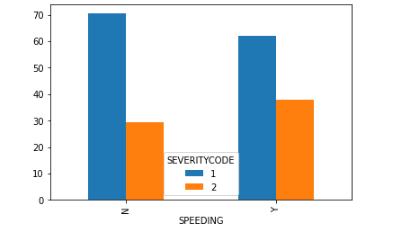
**3.6 Relationship between severity and Light condition:**

**There is a little increase in severity in dawn, dusk, and day light (chi square p<5%).**



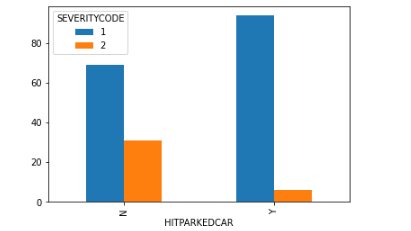
**3.7 Relationship between severity and Speeding:**

**It leads to more sever accidents (chi square p<5%).**



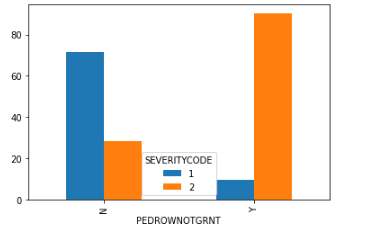
**3.8 Relationship between severity and Parked car:**

**If parked car existed in the accident, we can predict lower severity (chi square p<5%).**



**3.9 Relationship between severity and pedestrian right of way:**

**combined with higher severity (chi square p<5%).**



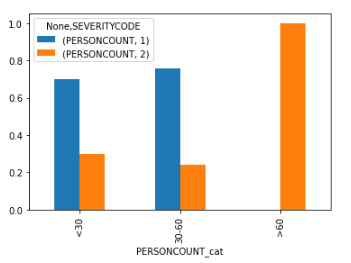
**3.10 Relationship between severity and Date & Time:**

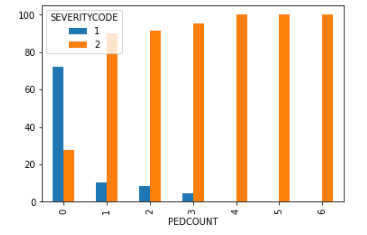
**There is an increase in severity between 5-8 am and 15-19 pm. (chi square p<5%). We notice also some variability regarding months and years which is statistically significant. (chi square p<5%). Contrary, Day of the month plays no role (chi square >5%). Years describe the past and not important to predict the future so I will not use it in my model.**

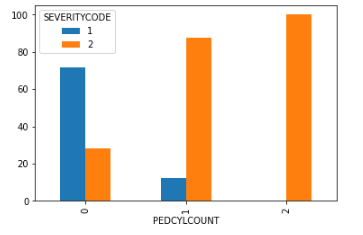
|  |  |
| --- | --- |
|  |  |
| **Hours** | **Months** |
|  |  |
| **Years** | **Days** |

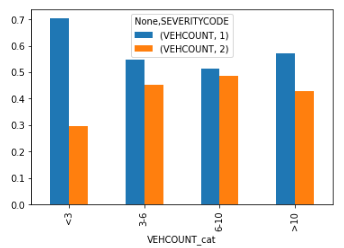
**3.11 Relationship between severity and Person Count & pedestrians & bicycles & vehicles:**

**We notice that the percentage of more severe accidents increases in upper classes for each feature (Mann Whitney U test, p<5%)**



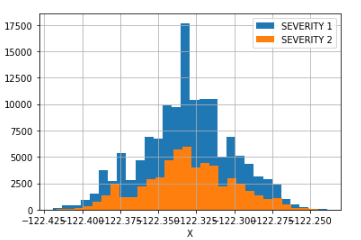


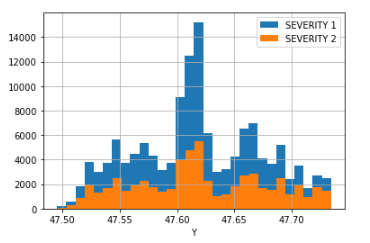




**3.12 Relationship between severity and Location:**

**We applied T test to detect if there is any statistically significant difference between groups. We found that P <5% for both X & Y.**





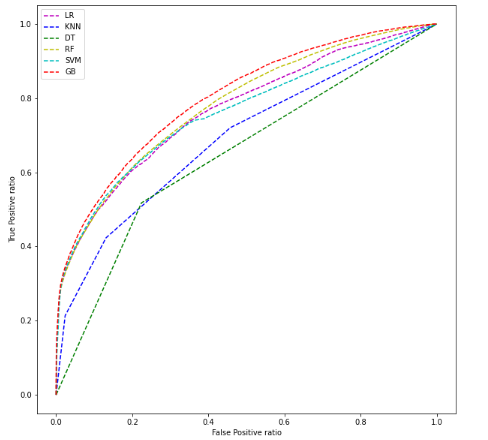
**4. Predictive Modeling:**

**There are two types of models, regression and classification, that can be used to predict severity. Here, we will use classification model because we have 2 discrete values for severity. Logistic regression, K-nearest neighbor, Decision tree, random forest, gradient boost models and a voting model were tuned and built.**

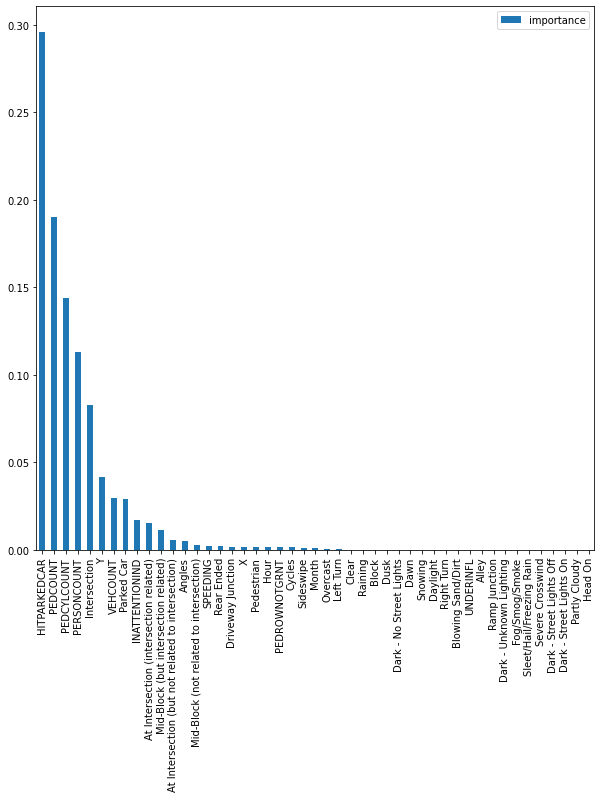
**Dataset is split to train (67%) and test set (33%). And various models are trained with good result as in table (1). gradient boost classifier achieved best result (highest accuracy & lowest log\_loss). Features, which is important in prediction process, is hitting parked car, number of persons involved (Person Count & pedestrians & bicycles), and intersection related accidents.**

**Table 1 models result**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **LR** | **KNN** | **DT** | **RF** | **GB** | **SVM** |
| **Accuracy** | **0.7847** | **0.7505** | **0.7110** | **0.7850** | **0.7911** | **0.7784** |
| **Log\_loss** | **0.4808** | **24.2034** | **9.9748** | **0.4828** | **0.4502** | **0.5179** |
| **Tp** | **6715** | **7581** | **10313** | **8751** | **7320** | **5460** |
| **Tn** | **43699** | **40638** | **35370** | **41681** | **43503** | **44550** |
| **Fp** | **1319** | **4380** | **9648** | **3337** | **1515** | **468** |
| **Fn** | **12510** | **11644** | **8912** | **10474** | **11905** | **13765** |



**Figure 1 ROC curve for models**



**Figure 2 Importance of features in GBC model**

**5. Conclusion:**

**In this study, I analyzed the relationship between severity of accidents and Features of the accident. I identified Location, Date, address type, junction type, inattention of driver, drugs or alcohol usage, weather condition, light condition, whether pedestrian right of way, speed, existing of parked car, where collusion occurred lane segment or crosswalk or intersection, type of collision, and count of persons, pedestrians, vehicle, and bicycles.**

**I built classification models with acceptable accuracy to predict accident severity. These models can be very useful for helping Government agencies, healthcare service provider, and Traveler in many ways.**