Coursera Capstone Project- Week 1/ Boaz Ben-Dov

**Car Accidents Severity Factors**

**Problem:**

A lot of effort and funding are invested every year all around the world to decrease car accidents severity outcomes. Considering every state has limited resources for dealing with the problem, it will make great importance to know where to focus our efforts to get the best results. I will try to answer this question by creating a model for classifying accidents severity and determine which factors effects the most.

**Data:**

I'll be using Seattle government collected data of 194,673 collisions between 01/01/2014 to 20/05/2020.

The independents variables can be divided to three categories:

1. External Factors- Weather, light
2. Drivers state- Under influence
3. Collison course- how many vehicles/pedestrians involved, what type of vehicles, where and how the car got hit.
4. Collison Location- Interchange, mid-block, specific location (prone to accidents).

The target variable is the Collison severity= property damage or injury

**Methodology:**

The car severity factors will be identified and quantified by using logistic regression model with target variable is high severity = 1 or low severity=0.

The logistic regression explanatory variables will be selected using **Recursive Feature Elimination (R.F.E)** algorithm. RFE works by searching for a subset of features by starting with all features in the training dataset and successfully removing features until the desired number remains.

Using **Synthetic Minority Oversampling Technique** (**SMOTE)** we will try to balance the data set- equalizing the severity category labels. SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.

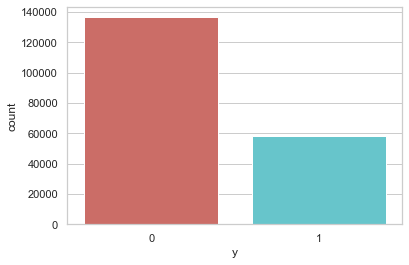
# After we have our logistic regression set, will use the p-value to point out the significance factors and because the data will be standardized we will use the variables coefficients to determine their importance.

# Will then discuss which factors are within authorities control and the way they increase accidents severity.

**Analysis:**

**Target Variable**

First, will explore the model target variable- Collison severity:



58,188

29.9%

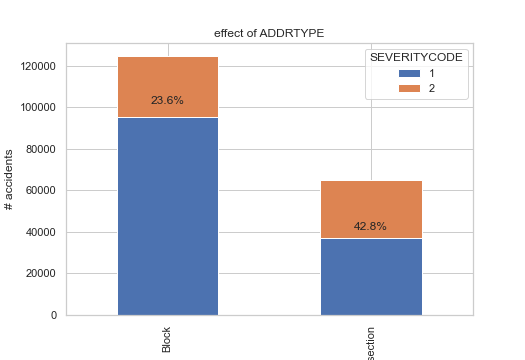
136,485

70.1%

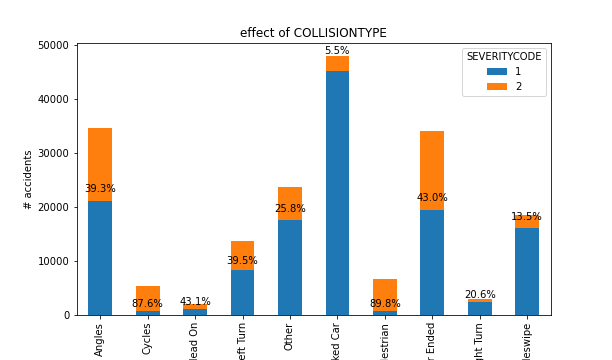
Clearly classification are not balanced. I'll address it later in the analysis.

**Explanatory variables**

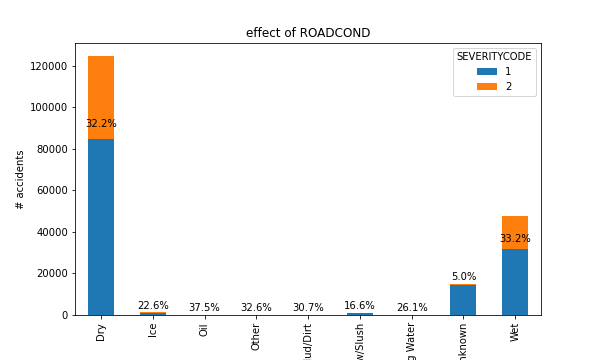
Let us get a feel about the effect of some of the variables by visualizing their relation to the target variable.



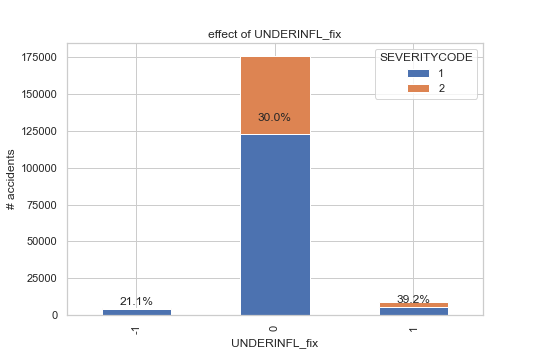
* Intersection are more dangerous than blocks



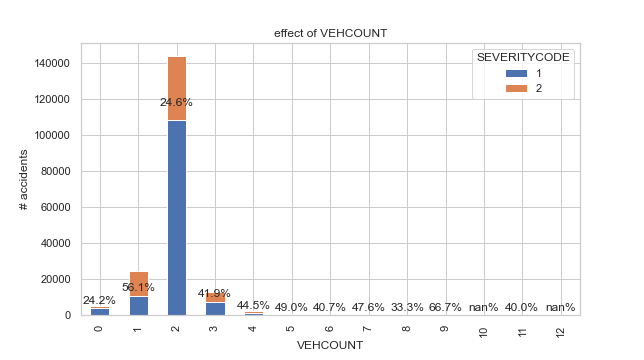
* The most sever collision happens, of course, when hitting a pedestrian or a cyclist and the least when hitting a parked car.
* Left turn is twice more dangerous than right turn.



* Surprisingly dry or wet road doesn't affect much Collison severity.



* As we expect the driving under influence increase collision severity.

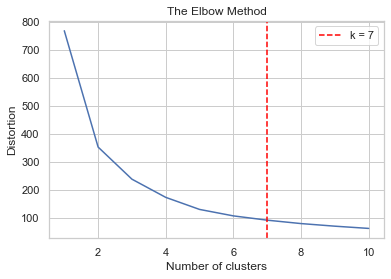


* When one vehicle is involved the severity is the highest (perhaps because this includes hitting a pedestrian or a cyclist).
* Two vehicles accidents are the most common and almost the least sever.
* Above two vehicles and the severity increases.

**Accident geographical location as explanatory variable:**

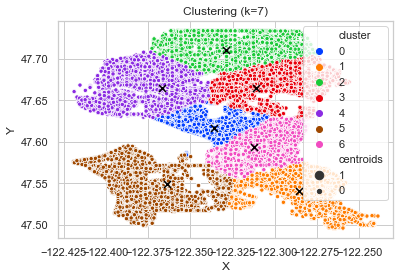
In order to understand if and how the geographic location of the accidents affects it severity, I decided to divide the accidents x-y coordinates into clusters and test the cluster effect on the accident severity. The clustering will be done using k-means.

To decide how many clusters will need, I've used the elbow method- an algorithm that help us to pick the point so that adding another cluster doesn't give much better modeling of the data- the curve of the graph:

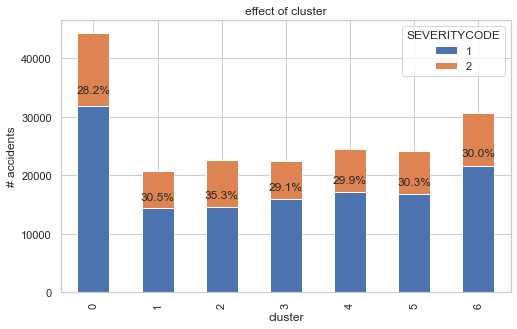


It is clear from the graph will need to use 7 clusters.

Running k-means on the accidents coordinated gives us this results:



We then examine the severity rate of each cluster:



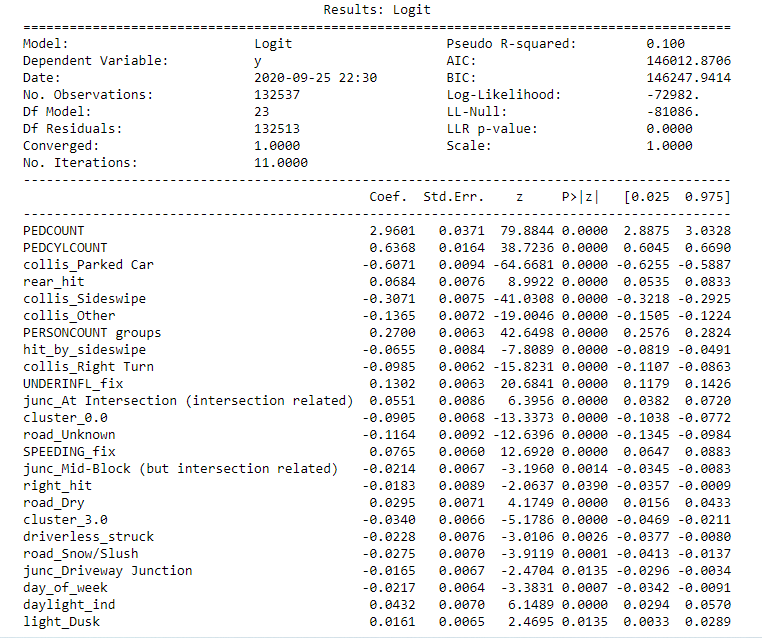
* Cluster 0 has the most accidents but also the lowest severity rate.
* Cluster 2 stands out with the highest severity rate.

**Results:**

The RFE picked up the flowing variables:

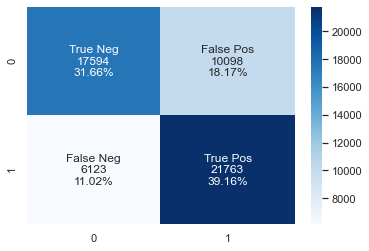
|  |  |  |
| --- | --- | --- |
| rear\_hit | junc\_Mid-Block (not related to intersection) | PERSONCOUNT |
| left\_hit | wthr\_Unknown | PEDCOUNT |
| right\_hit | road\_Dry | PEDCYLCOUNT |
| hit\_by\_angle | road\_Snow/Slush | VEHCOUNT |
| hit\_by\_sideswipe | road\_Unknown | INTCHNG\_IND |
| daylight\_ind | road\_nan | collis\_Angles |
| light\_Dark - Street Lights On | UNDERINFL\_fix | collis\_Cycles |
| light\_Daylight | SPEEDING\_fix | collis\_Left Turn |
| light\_Dusk | driverless\_struck | collis\_Other |
| light\_Unknown | motor\_struck | collis\_Parked Car |
| day\_of\_week | pedalcyclist\_struck | collis\_Pedestrian |
| PERSONCOUNT groups | idk\_struck | collis\_Right Turn |
| VEHCOUNT groups | struck\_motor | collis\_Sideswipe |
| cluster\_0.0 | struck\_pedalcyclist | junc\_At Intersection (intersection related) |
| cluster\_3.0 | struck\_self | junc\_Driveway Junction |
|  | front\_hit | junc\_Mid-Block (but intersection related) |

After keeping only significant variables (p-value<0.05), we are left with the following regression:



The regression produce the following results:

Confusion matrix:



Classification report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| 0 | 0.74 | 0.64 | 0.68 | 27,692 |
| 1 | 0.68 | 0.78 | 0.73 | 27,886 |

Roc Curve:



* The roc curve indicates this is a fair regression

**Discussion:**

Though the regression accuracy measures are not very good, we still can try to conclude from the variables coefficients which factors effect accidents severity more and how.

Because all the variables are all standardized the coefficients are a good measure of the variable effect. Let's sort the variables by the absolute value of the coefficients and try to describe their effect:

|  |  |  |  |
| --- | --- | --- | --- |
| variable | coef | abs(coef) | Description |
| PEDCOUNT | 2.9601 | 2.9601 | The more pedestrians are involved the severity rate increase |
| PEDCYLCOUNT | 0.6368 | 0.6368 | The more bicycles are involved the severity rate increase |
| collis\_Parked Car | -0.6071 | 0.6071 | Collision with parked car decrease the severity rate |
| collis\_Sideswipe | -0.3071 | 0.3071 | Sideswipe collision decrease severity rate |
| PERSONCOUNT groups | 0.27 | 0.27 | The more people are involved in the collision, the severity rate increase |
| collis\_Other | -0.1365 | 0.1365 | "Other" collision decrease severity rate- further understanding of this variable is needed |
| UNDERINFL\_fix | 0.1302 | 0.1302 | Collision under influence of drugs and alcohol increase the severity rate |
| road\_Unknown | -0.1164 | 0.1164 | unknown road condition decrease severity rate- further understanding of this variable is needed |
| collis\_Right Turn | -0.0985 | 0.0985 | Collision while making right turn decrease the severity rate |
| cluster\_0.0 | -0.0905 | 0.0905 | Collision in geographic cluster 0 increase the severity rate |
| SPEEDING\_fix | 0.0765 | 0.0765 | Collision while speeding increase the severity rate |
| rear\_hit | 0.0684 | 0.0684 | Collision in rear increase severity rate |
| hit\_by\_sideswipe | -0.0655 | 0.0655 | Collision hit by sideswipe decrease severity rate |
| junc\_At Intersection (intersection related) | 0.0551 | 0.0551 | Collision at intersection increase the severity rate |
| daylight\_ind | 0.0432 | 0.0432 | Collision when daylight increase the severity rate |
| cluster\_3.0 | -0.034 | 0.034 | Collision in geographic cluster 3 decrease the severity rate |
| road\_Dry | 0.0295 | 0.0295 | Collision on a dry road increase the severity rate |
| road\_Snow/Slush | -0.0275 | 0.0275 | Collision on a snowed road decrease the severity rate |
| driverless\_struck | -0.0228 | 0.0228 | Driverless collision decrease the severity rate |
| day\_of\_week | -0.0217 | 0.0217 | As value of day of week higher (1-7) the severity rate decrease |
| junc\_Mid-Block (but intersection related) | -0.0214 | 0.0214 | Collision on mid-block decrease severity rate |
| right\_hit | -0.0183 | 0.0183 | Collision on right side decrease severity rate |
| junc\_Driveway Junction | -0.0165 | 0.0165 | Collision on driveway decrease severity rate |
| light\_Dusk | 0.0161 | 0.0161 | Collision when dusk increase severity rate |

So where should authorities focus their efforts in order to decrease accidents severity:

* It comes with no surprise that what effects accidents severity the most is when pedestrians and bicycles are involved. Authorities should probably invest more efforts finding solutions for keeping them safe (perhaps more bicycle roads and lowering speeding limits within populated areas).
* Moderate effect on severity rate can be achieved by decreasing speeding and driving under influence.
* Other variables either shows small effect or behind the control of state authorities in my opinion.

**Conclusions:**

Unfortunately this project didn't deliver the best results measure wise, so it's problematic to try and draw any hard conclusions.

More data on the drivers (like age, gender, driving experience, driving record) and on the vehicles (like safety rate and safety features) could have help us to build a better model.

Having said that, judging by this model results, state authorities should focus their efforts on keeping pedestrians and cyclists safe if they want to decrease their accidents severity rate.

Though the geographic clustering didn't produce much effect on the severity rate, I think it will be worth the while to create a more comprehensive geographic model to try to understand what factors effect accidents severity rate the most on each area.