Collision in Seattle

Table of Contents

[Introduction 1](#_Toc54126639)

[Data Preparation 2](#_Toc54126640)

[Data Cleaning 2](#_Toc54126641)

[Data after removal 3](#_Toc54126642)

[Handle Null Value 3](#_Toc54126643)

[Methodology 3](#_Toc54126644)

[Correct Data Types 4](#_Toc54126645)

[Data to Predict 4](#_Toc54126646)

[Data Exploration 5](#_Toc54126647)

[Convert Categorical Data to Numeric 13](#_Toc54126648)

[Spit Testing and Training Data 14](#_Toc54126649)

[Data Analysis 14](#_Toc54126650)

[Pearson Correlation 14](#_Toc54126651)

[Decision Tree 14](#_Toc54126652)

[K-Nearest Neighbor (KNN) 15](#_Toc54126653)

[Logistic Regression 15](#_Toc54126654)

[Support Vector Machine (SVM) 16](#_Toc54126655)

[Conclusion 17](#_Toc54126656)

# Introduction

There are many accidents occur in Seattle. Each incident has different severity level, such as prop damage, injury, serious injury, or even fatality. Since we have raw data for each accident, one way to reduce it is to learn from the accident that was occurred. To see what behavioral reflects the severity of an accident. This can help us understand the nature of the accident and we can precede preventive action to change on the properties that lead to high severity accident.

# Data Preparation

## Data Cleaning

There are 37 attribute and 194,673 rows in the raw data that we can use to learn. However, some data is not useful for the analysis which is:

1. no meaning and duplicated data from the data set (a key data, description data)
   1. 'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS', 'LOCATION', 'EXCEPTRSNDESC', 'SEVERITYDESC', 'SDOTCOLNUM', 'ST\_COLDESC', 'SEGLANEKEY', 'CROSSWALKKEY', 'INTKEY', 'SEVERITYCODE.1', 'INCDTTM', 'INCDATE', 'SDOT\_COLDESC', ‘INCDTTM’
2. data contain too many null value which is easily lead to incorrect prediction (more than 50% of the data)
   1. 'EXCEPTRSNCODE', 'INATTENTIONIND', 'PEDROWNOTGRNT', 'SPEEDING'

Data that is interesting to be a feature to solve the problem is listed as the following.

1. ADDRTYPE - Collision address type
   1. Sample data = {Alley, Block, Intersection}
2. COLLISIONTYPE - Collision type
   1. Sample data = {Parked car, Angles, Rear Ended, Sidewipe, Left Turn, RightTurn, Pedestrian, Cycles, Head On, Other}
3. PERSONCOUNT - The total number of people involved in the collision
4. PEDCOUNT - The number of pedestrians involved in the collision
5. PEDCYLCOUNT - The number of bicycles involved in the collision
6. VEHCOUNT - The number of vehicles involved in the collision
7. JUNCTIONTYPE - Category of junction at which collision took place
   1. Sample data = {'At Intersection (intersection related)', 'Mid-Block (not related to intersection)', 'Driveway Junction', 'Mid-Block (but intersection related)', 'At Intersection (but not related to intersection)', 'Unknown', 'Ramp Junction'}
8. SDOT\_COLCODE - A code given to the collision by SDOT
   1. Sample data = 11 means MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END AT ANGLE, 16 means MOTOR VEHICLE STRUCK MOTOR VEHICLE, LEFT SIDE SIDESWIPE
9. UNDERINFL - Whether or not a driver involved was under the influence of drugs or alcohol.
10. WEATHER - A description of the weather conditions during the time of the collision
    1. Sample data = {Clear, Raining, Overcast, Snowing, Fog/Smog/Smoke, Sleet/Hail/Freezing Rain, Blowing Sand/Dirt, Severe Crosswind, Partly Cloudy, Other, Unknown}
11. ROADCOND - The condition of the road during the collision
    1. Sample data = {Dry, Wet, Unknown, Ice, Snow/Slush, Standing Water, Sand/Mud/Dirt, Oil, Other}
12. LIGHTCOND - The light conditions during the collision
    1. Sample data = {Daylight, Dark - Street Lights On, Dark - No Street Lights, Unknown, Dusk, Dawn, Dark - Street Lights Off, Other, Dark - Unknown Lighting}
13. ST\_COLCODE - A code provided by the state that describes the collision
    1. Sample data = 0 means 'Vehicle Going Straight Hits Pedestrian', 11 means 'From Same Direction -Both Going Straight-Both Moving- Sideswipe'
14. HITPARKEDCAR - Whether or not the collision involved hitting a parked car. (Y/N)

These columns have below data type.

|  |
| --- |
| SEVERITYCODE int64  ADDRTYPE object  COLLISIONTYPE object  PERSONCOUNT int64  PEDCOUNT int64  PEDCYLCOUNT int64  VEHCOUNT int64  JUNCTIONTYPE object  SDOT\_COLCODE int64  UNDERINFL object  WEATHER object  ROADCOND object  LIGHTCOND object  ST\_COLCODE object  HITPARKEDCAR object |

### Data after removal

There are 15 columns of data and 194,673 rows

## Handle Null Value

The data has null value as below.

|  |
| --- |
| SEVERITYCODE 0  ADDRTYPE 1926  COLLISIONTYPE 4904  PERSONCOUNT 0  PEDCOUNT 0  PEDCYLCOUNT 0  VEHCOUNT 0  JUNCTIONTYPE 6329  SDOT\_COLCODE 0  UNDERINFL 4884  WEATHER 5081  ROADCOND 5012  LIGHTCOND 5170  ST\_COLCODE 18  HITPARKEDCAR 0 |

### Methodology

|  |  |
| --- | --- |
| Data field | Method |
| ST\_COLCODE | Replace blank value (‘ ‘) as NaN and replace NaN with maximum frequency data |
| JUNCTIONTYPE | Replace ‘Unknown’ value as NaN and replace NaN with maximum frequency data |
| UNDERINFL | Replace ‘Y’ as 1 and ‘N’ as 0 and replace NaN with maximum frequency data |
| HITPARKEDCAR | Replace ‘Y’ as 1 and ‘N’ as 0 |
| ADDRTYPE | Replace NaN with maximum frequency data |
| COLLISIONTYPE | Replace NaN with maximum frequency data |
| WEATHER | Replace NaN with maximum frequency data |
| ROADCOND | Replace NaN with maximum frequency data |
| LIGHTCOND | Replace NaN with maximum frequency data |

## Correct Data Types

Change ST\_COLCODE, UNDERINFL, and HITPARKEDCAR to integer. After changing, the data types is as below.

|  |
| --- |
| SEVERITYCODE int64  ADDRTYPE object  COLLISIONTYPE object  PERSONCOUNT int64  PEDCOUNT int64  PEDCYLCOUNT int64  VEHCOUNT int64  JUNCTIONTYPE object  SDOT\_COLCODE int64  UNDERINFL int64  WEATHER object  ROADCOND object  LIGHTCOND object  ST\_COLCODE int64  HITPARKEDCAR int64 |

## Data to Predict

Data to predict is SEVERITYCODE. The data in this data set contain only 2 values as below.

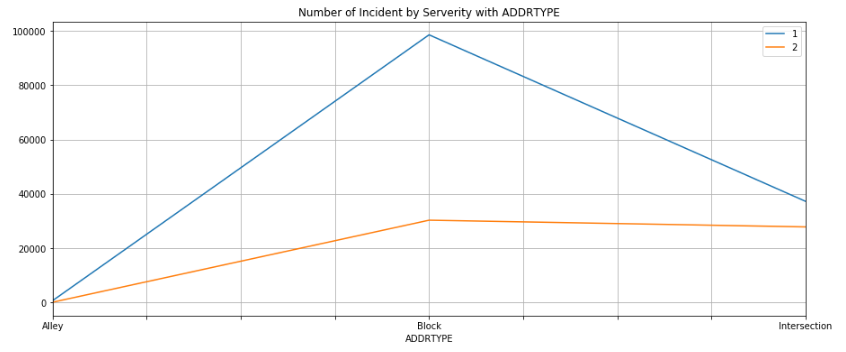
|  |  |  |
| --- | --- | --- |
| Severity Code | Description | Amount of data |
| 1 | Prop damage | 136,485 |
| 2 | Injury | 58,188 |



## Data Exploration

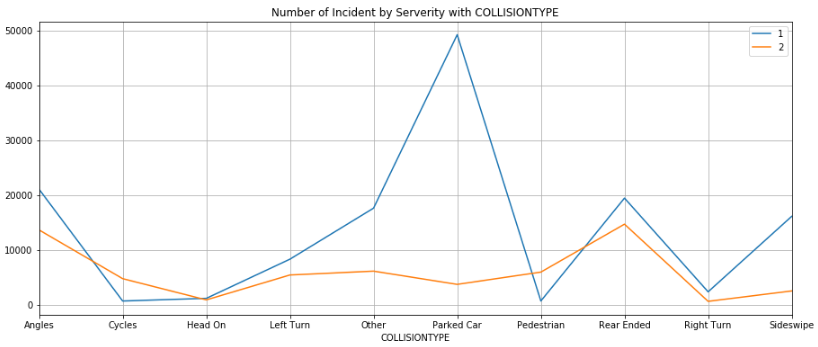
ADDRTYPE data group by severity code

|  |
| --- |
| ADDRTYPE SEVERITYCODE COUNT  0 Alley 1 669  1 Alley 2 82  2 Block 1 98565  3 Block 2 30287  4 Intersection 1 37251  5 Intersection 2 27819 |



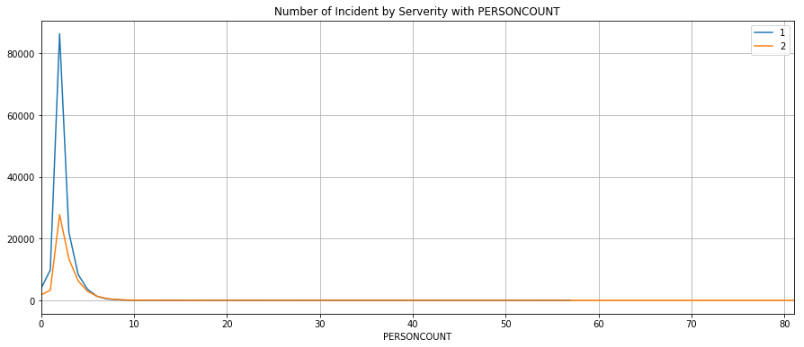
COLLISIONTYPE data group by severity code

|  |
| --- |
| COLLISIONTYPE SEVERITYCODE COUNT  0 Angles 1 21050  1 Angles 2 13624  2 Cycles 1 671  3 Cycles 2 4744  4 Head On 1 1152  5 Head On 2 872  6 Left Turn 1 8292  7 Left Turn 2 5411  8 Other 1 17591  9 Other 2 6112  10 Parked Car 1 49188  11 Parked Car 2 3703  12 Pedestrian 1 672  13 Pedestrian 2 5936  14 Rear Ended 1 19419  15 Rear Ended 2 14671  16 Right Turn 1 2347  17 Right Turn 2 609  18 Sideswipe 1 16103  19 Sideswipe 2 2506 |



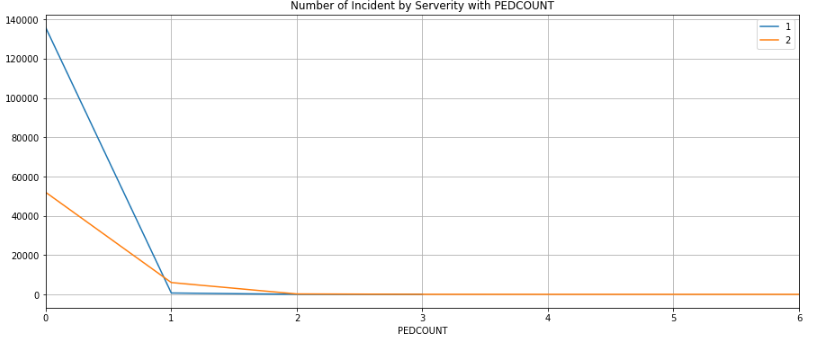
PERSONCOUNT data group by severity code

|  |
| --- |
| PERSONCOUNT SEVERITYCODE COUNT  0 0 1 3782  1 0 2 1762  2 1 1 9858  3 1 2 3296  4 2 1 86420  5 2 2 27811  6 3 1 22092  7 3 2 13461  8 4 1 8365  9 4 2 6295  10 5 1 3615  11 5 2 2969  12 6 1 1345  13 6 2 1357  14 7 1 494  15 7 2 637  16 8 1 249  17 8 2 284  18 9 1 87  19 9 2 129  20 10 1 54  21 10 2 74  22 11 1 23  23 11 2 33  24 12 1 13  25 12 2 20  26 13 1 9  27 13 2 12  28 14 1 12  29 14 2 7  .. ... ... ...  61 32 1 2  62 32 2 1  63 34 1 1  64 34 2 2  65 35 1 1  66 36 1 2  67 37 1 2  68 37 2 1  69 39 2 1  70 41 1 1  71 43 1 1  72 44 1 6  73 47 1 3  74 48 2 1  75 53 1 1  76 54 2 1  77 57 1 1  78 81 2 1 |



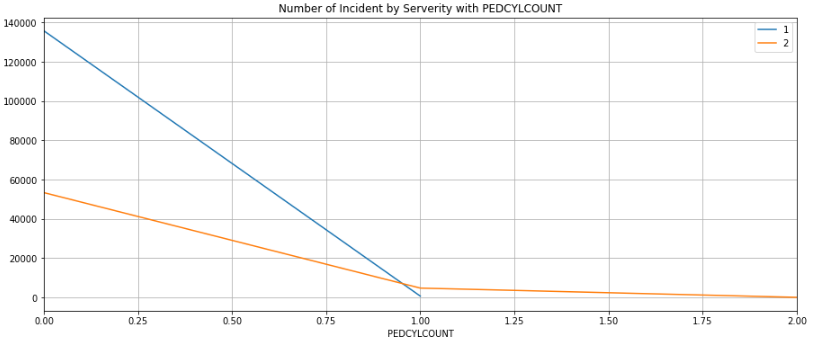
PEDCOUNT data group by severity code

|  |
| --- |
| PEDCOUNT SEVERITYCODE COUNT  0 0 1 135787  1 0 2 51947  2 1 1 678  3 1 2 6007  4 2 1 19  5 2 2 207  6 3 1 1  7 3 2 21  8 4 2 4  9 5 2 1  10 6 2 1 |



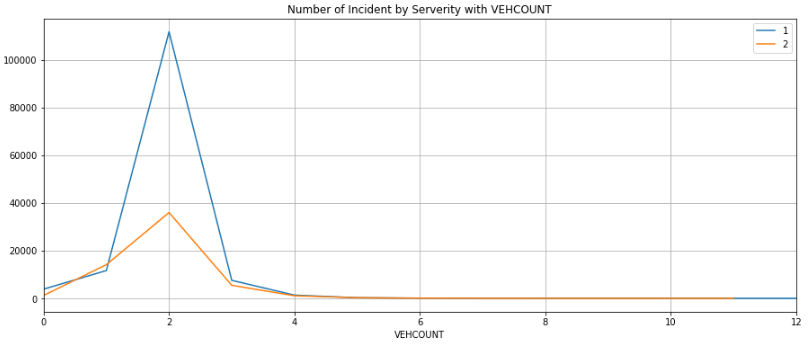
PEDCYLCOUNT data group by severity code

|  |
| --- |
| PEDCYLCOUNT SEVERITYCODE COUNT  0 0 1 135806  1 0 2 53383  2 1 1 679  3 1 2 4762  4 2 2 43 |



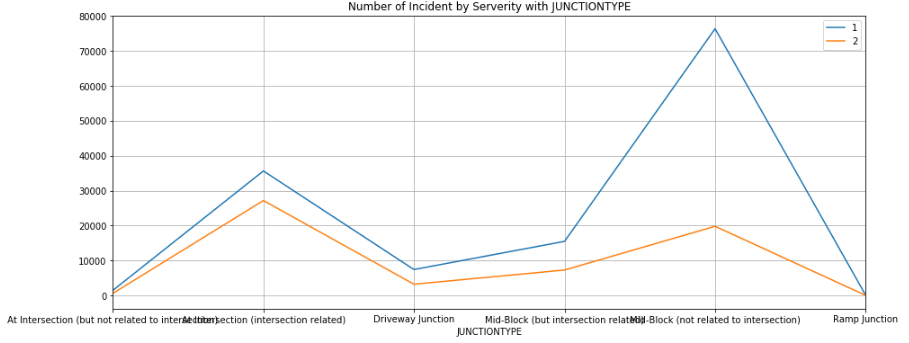
VEHCOUNT data group by severity code

|  |
| --- |
| VEHCOUNT SEVERITYCODE COUNT  0 0 1 3858  1 0 2 1227  2 1 1 11643  3 1 2 14105  4 2 1 111701  5 2 2 35949  6 3 1 7540  7 3 2 5470  8 4 1 1348  9 4 2 1078  10 5 1 268  11 5 2 261  12 6 1 86  13 6 2 60  14 7 1 24  15 7 2 22  16 8 1 10  17 8 2 5  18 9 1 3  19 9 2 6  20 10 2 2  21 11 1 3  22 11 2 3  23 12 1 1 |



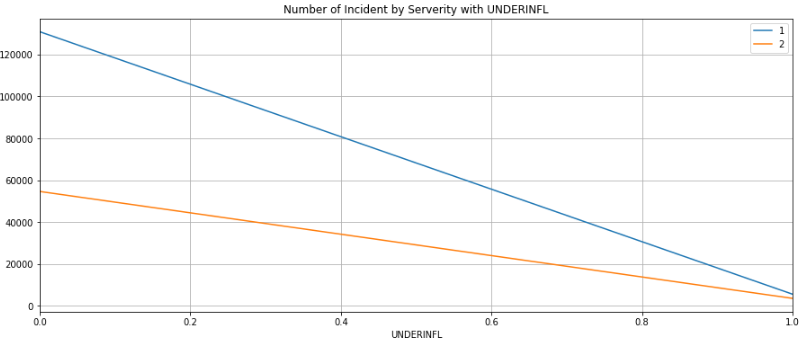
JUNCTIONTYPE data group by severity code

|  |
| --- |
| JUNCTIONTYPE SEVERITYCODE COUNT  0 At Intersection (but not related to intersection) 1 1475  1 At Intersection (but not related to intersection) 2 623  2 At Intersection (intersection related) 1 35636  3 At Intersection (intersection related) 2 27174  4 Driveway Junction 1 7437  5 Driveway Junction 2 3234  6 Mid-Block (but intersection related) 1 15493  7 Mid-Block (but intersection related) 2 7297  8 Mid-Block (not related to intersection) 1 76332  9 Mid-Block (not related to intersection) 2 19806  10 Ramp Junction 1 112  11 Ramp Junction 2 54 |



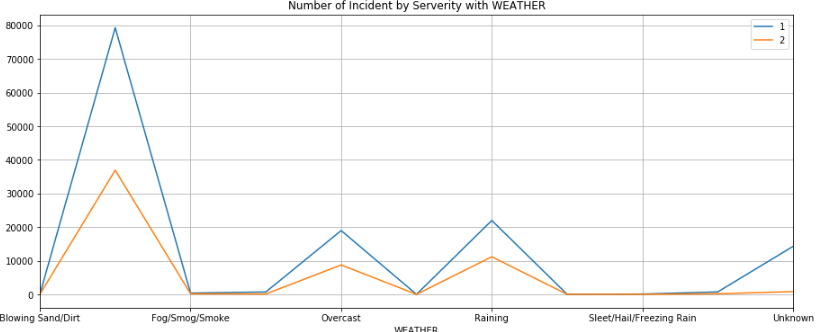
UNDERINFL data group by severity code

|  |
| --- |
| UNDERINFL SEVERITYCODE COUNT  0 0 1 130926  1 0 2 54626  2 1 1 5559  3 1 2 3562 |



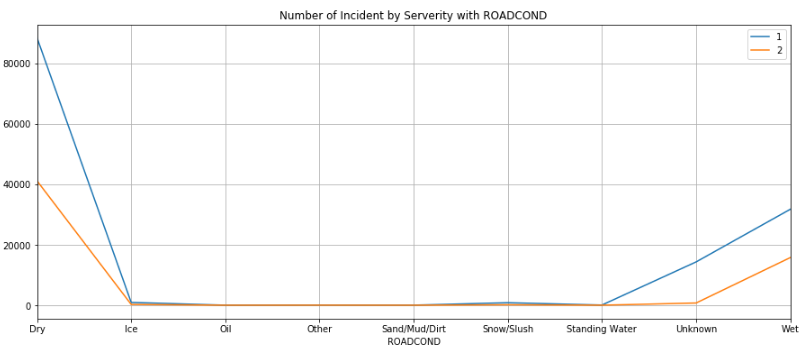
WEATHER data group by severity code

|  |
| --- |
| WEATHER SEVERITYCODE COUNT  0 Blowing Sand/Dirt 1 41  1 Blowing Sand/Dirt 2 15  2 Clear 1 79292  3 Clear 2 36924  4 Fog/Smog/Smoke 1 382  5 Fog/Smog/Smoke 2 187  6 Other 1 716  7 Other 2 116  8 Overcast 1 18969  9 Overcast 2 8745  10 Partly Cloudy 1 2  11 Partly Cloudy 2 3  12 Raining 1 21969  13 Raining 2 11176  14 Severe Crosswind 1 18  15 Severe Crosswind 2 7  16 Sleet/Hail/Freezing Rain 1 85  17 Sleet/Hail/Freezing Rain 2 28  18 Snowing 1 736  19 Snowing 2 171  20 Unknown 1 14275  21 Unknown 2 816 |



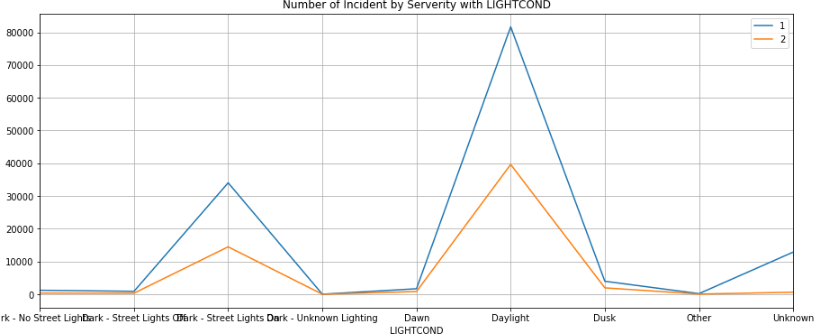
ROADCOND data group by severity code

|  |
| --- |
| ROADCOND SEVERITYCODE COUNT  0 Dry 1 88398  1 Dry 2 41124  2 Ice 1 936  3 Ice 2 273  4 Oil 1 40  5 Oil 2 24  6 Other 1 89  7 Other 2 43  8 Sand/Mud/Dirt 1 52  9 Sand/Mud/Dirt 2 23  10 Snow/Slush 1 837  11 Snow/Slush 2 167  12 Standing Water 1 85  13 Standing Water 2 30  14 Unknown 1 14329  15 Unknown 2 749  16 Wet 1 31719  17 Wet 2 15755 |



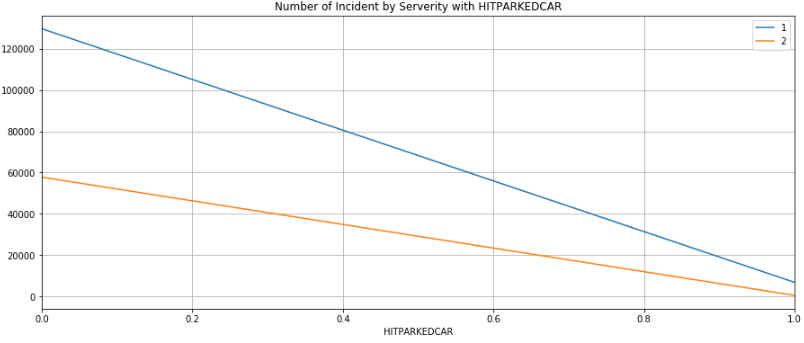
LIGHTCOND data group by severity code

|  |
| --- |
| LIGHTCOND SEVERITYCODE COUNT  0 Dark - No Street Lights 1 1203  1 Dark - No Street Lights 2 334  2 Dark - Street Lights Off 1 883  3 Dark - Street Lights Off 2 316  4 Dark - Street Lights On 1 34032  5 Dark - Street Lights On 2 14475  6 Dark - Unknown Lighting 1 7  7 Dark - Unknown Lighting 2 4  8 Dawn 1 1678  9 Dawn 2 824  10 Daylight 1 81673  11 Daylight 2 39634  12 Dusk 1 3958  13 Dusk 2 1944  14 Other 1 183  15 Other 2 52  16 Unknown 1 12868  17 Unknown 2 605 |



HITPARKEDCAR data group by severity code

|  |
| --- |
| HITPARKEDCAR SEVERITYCODE COUNT  0 0 1 129717  1 0 2 57740  2 1 1 6768  3 1 2 448 |



## Convert Categorical Data to Numeric

Since many libraries do not support categorical data, we have to convert them to Numeric value. Then normalize it.

## Spit Testing and Training Data

* X = 'ADDRTYPE', 'COLLISIONTYPE', 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'JUNCTIONTYPE', 'SDOT\_COLCODE', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'ST\_COLCODE', 'HITPARKEDCAR'
* y = ‘SEVERITYCODE’

|  |  |  |
| --- | --- | --- |
| Data | Percentage of all data | Amount of data |
| Training | 70% | 136,271 |
| Testing | 30% | 58,402 |

# Data Analysis

## Pearson Correlation

The Pearson Correlation measures the linear dependence between two variables X and Y.

The resulting coefficient is a value between -1 and 1 inclusive, where:

* 1: Total positive linear correlation.
* 0: No linear correlation, the two variables most likely do not affect each other.
* -1: Total negative linear correlation.

Result of the correlation with SEVERITYCODE shows as the following

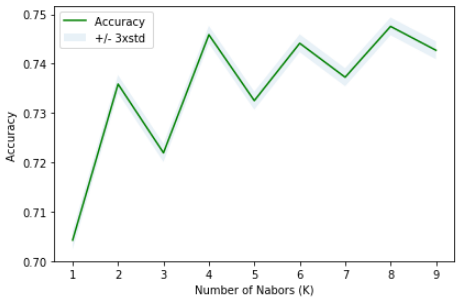
|  |  |
| --- | --- |
| Data | Pearson Correlation with SEVERITYCODE |
| SEVERITYCODE | 1 |
| PERSONCOUNT | 0.130949 |
| PEDCOUNT | 0.246338 |
| PEDCYLCOUNT | 0.214218 |
| VEHCOUNT | -0.054686 |
| SDOT\_COLCODE | 0.188905 |
| UNDERINFL | 0.044377 |
| ST\_COLCODE | -0.165233 |
| HITPARKEDCAR | -0.101498 |

There is no data that nearly -1 or 1 with SEVERITYCODE, all of them nearly to 0. Then there is no linear correlation to SEVERITYCODE. Linear regression should not be a good method to use to predict the severity code.

## Decision Tree

* Max depth = 5
* Accuracy (the fraction of correctly classified samples) = 0.7537584329303791

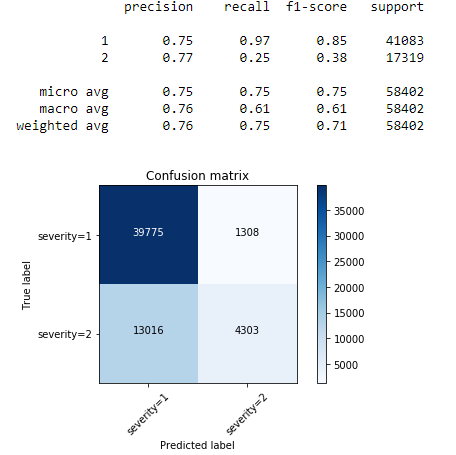
## K-Nearest Neighbor (KNN)

* From experiments, K is 1 to 10, the best K is 8.  
  
* Accuracy (the fraction of correctly classified samples) = 0.7475771377692545

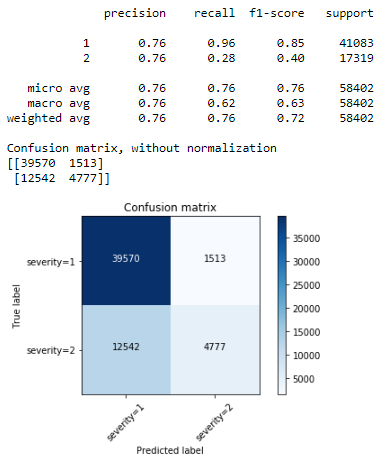
## Logistic Regression

* Use Solver Liblinear
* Accuracy
  + Jaccard similarity score = 0.7547344269031883

Note that this is equal to the fraction of correctly classified samples.

* + Confusion matrix (F1 average) = 0.7173683103496501  
    

## Support Vector Machine (SVM)

* Kernel ‘rbf’
* Accuracy
  + The fraction of correctly classified samples = 0.7593404335467964
  + Confusion matrix (F1 average) = 0.7173683103496501  
    

# Conclusion

Collision in Seattle happens with many properties. In this study we use properties which are collision address type, collision type, number of people involved in the collision, number of pedestrians, number of bicycles, number of vehicles, category of junction, collision code by SDOT, driver was under influence of drug or alcohol, weather, road condition, light condition, a code that provided by the state, and whether or not the collision involved hitting a parked car, to predict severity level of the collision. The severity level can be prop damage, injury, serious injury, or even fatality.

From the source data, first, the study removed the data which has a duplicated meaning, and data that has null value more than 50%. Second, handle null value with the data that has the highest frequency. Third, change data type to a proper one. Fourth, did the data normalization. Finally, split this data to training set (70% of the data) and testing set (30% of the data).

From the study, Decision tree algorithm, K-Nearest Neighbor algorithm (KNN), Logistic Regression, and Support Vector Machine algorithm (SVM), shows similar ability to predict the severity level. Decision tree and SVM has accuracy score 76%, where KNN and Logistic Regression have accuracy score 75%.