**Data Science in Car Accident Severity**

1. **Introduction/Business Problem***.*

Car accidents are a huge problem in our world. This project aims in analyzing "car accident severity” in terms of human fatality, traffic delay, property damage, or any other chance of fatality. In order to reduce car collisions in a community, a data science model must be trained to predict the severity of an accident. Taking consideration of the current weather, road and visibility conditions. When conditions are bad, this model will alert drivers to remind them to be more careful.

1. **Data section**

The dataset for this project is taken from a shared data link which is collected from Seattle SPOT Traffic Management Division. This is the shared data for Seattle city. The dataset is in the form of .CSV file. This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. The target label for the dataset is severity, which describes the fatality of an accident. The shared data has unbalanced labels. This dataset is updated weekly and is from 2004 to present. It contains information such as severity code, address type, location, collision type, weather, road condition, speeding, etc.,. There are 37 attributes in this dataset.

This Project for everyone who really care about the traffic records, especially in the transportation department.

This model is to improve the predictability of the accident severity and to reduce accidents in the future.

The result helps SHSP, DMV, stakeholders, insurance company, car manufacturers, and partnerships to allocate budget for education and enforcement to act on the result in order to achieve the goal of minimizing fatal/injury car crash.

The link for the dataset: <https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv>

The link for the Metadata of the dataset: <https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf>

**Some of the Metadata given with the dataset:**

The target Data to be predicted under (**SEVERITYCODE** 1-prop damage 2-injury) label.

Other important variables include:

* **ADDRTYPE**: Collision address type: Alley, Block, Intersection
* **LOCATION**: Description of the general location of the collision
* **PERSONCOUNT**: The total number of people involved in the collision helps to identify severity involved
* **PEDCOUNT**: The number of pedestrians involved in the collision helps to identify severity involved
* **PEDCYLCOUNT**: The number of bicycles involved in the collision helps to identify severity involved
* **VEHCOUNT**: The number of vehicles involved in the collision helps to identify severity involved
* **JUNCTIONTYPE**: Category of junction at which collision took place helps to identify where most collisions occur
* **WEATHER**: A description of the weather conditions during the time of the collision
* **ROADCOND**: The condition of the road during the collision
* **LIGHTCOND**: The light conditions during the collision
* **SPEEDING**: Whether speeding was a factor in the collision (Y/N)
* **HITPARKEDCAR**: Whether the collision involved hitting a parked car

1. **Methodology section** 
   1. **Data Analysis:**

We have unbalanced dataset hence we must balance it. We should extract and convert the dataset into a proper format.

Machine generated alternative text:
SEVERITYCODE 
2 
2 
-122323148 
-122347294 
-122334540 
-122334803 
-122306426 
Y 
47703140 
47647172 
47607871 
47604803 
47545739 
OBJECTID 
3 
5 
INCKEY 
1307 
26700 
1144 
17700 
COLDETKEY 
1307 
52200 
26700 
1144 
17700 
REPORTNO 
3502005 
2607959 
1482393 
3503937 
1807429 
STATUS 
Matched 
Matched 
Matched 
Matched 
Matched 
ADDRTYPE 
Intersection 
Block 
Block 
Block 
Intersection 
INTKEY 
374750 
343870 
ROADCOND 
Dry 
Dry 
LIGHTCOND 
Daylight 
Dark 
Street 
Lights On 
Daylight 
Daylight 
Daylight 

After dropping the unwanted columns and unknown numbers (NaN) the data types of the new columns in our data frame

Machine generated alternative text:
SEVER 1 TYCODE 
ADDRTYPE 
LOCATION 
EXCEPTRSNCODE 
EXCEPTRSNDEsc 
SEVERITYCODE . 1 
SEVERITYDEsc 
COLLISIONTYPE 
PER s ONCOUNT 
PE DCOUNT 
PE DCYLCOUNT 
VEHCOUNT 
INCDATE 
INCDTTU 
JUNCTIONTYPE 
INATTENTIONIND 
UNDERINE'L 
ROAD COND 
LIGHTCOND 
SPEEDING 
HIT PARKEDCAR 
dtype : obje ct 
int 64 
oaté4 
oaté4 
obj e ct 
obj e ct 
obj e ct 
obj e ct 
int 64 
obj e ct 
obj e ct 
int 64 
int 64 
int 64 
int 64 
obj e ct 
obj e ct 
obj e ct 
obj e ct 
obj e ct 
obj e ct 
obj e ct 
obj e ct 
obj e ct 
obj e ct 

**Balancing the Dataset**

Our target variable SEVERITYCODE is only 42% balanced. In fact, severitycode in class 1 is nearly three times the size of class 2.

Machine generated alternative text:
Number of Severity based on code 
120000 - 
100000 - 
80000 - 
40000 - 

Calculating the total number of car accidents under different situations

|  |  |
| --- | --- |
| Machine generated alternative text: index  Clear  Raining  Overcast  un known  Snowing  FogJSmogJSmoke  SleeVHail/Freezjng Rain  Blowing SandJDir1  Severe Crosswind  Partly Cloudy  WEATHER  108959  32015  27136  13893 |  |

|  |  |
| --- | --- |
| Machine generated alternative text: index  Dry  Unknown  Snow/Slush  Standing Water  SandJMudJDlr1  ROADCOND  122076  13839  1177 |  |

|  |  |
| --- | --- |
| Machine generated alternative text: index  Daylight  Dark - Street Lights On  Unknown  Dusk  Dawn  Dark - No Street Lights  Dark - Street Lights Off  Dark - Unknown Lighting  LIGHTCOND  1135A9  47314  12432  5775  1451  1152 |  |

|  |
| --- |
| Machine generated alternative text: index  N NORTHGATE WAY BET%'EEN MERIDIAN AVE N AND COR„  AURORAAVEN BET%'EEN N 117TH PL AND N 125TH ST  6TH AVE AND JAMES ST  AURORAAVE N N 130TH ST AND N 135TH ST  RAINIER AVE S BET%'EEN S BAWIEW ST AND S MCCLE„  WEST SEATTLE BR EB BET'.A'EEN ALASKAN VI NB  AURORA BR BET%'EEN RAYE ST AND BRIDGE WAY N  1ST AVE BET%'EEN BLANCHARD ST AND BELL ST  5TH AVE AND SPRING ST  RAINIER AVE S S HENDERSON ST AND S  LOCATION  212 |
|  |

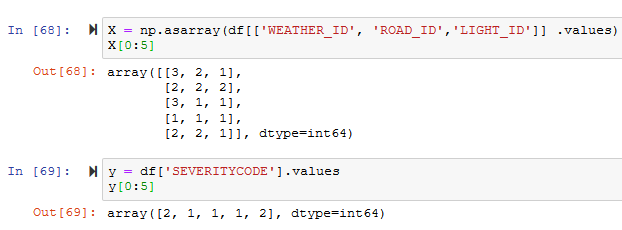
Machine generated alternative text:
Magnolia 
West Seattle 
tri. • 
Queen Anke 
Delridge 
uni rsity District= - 
9 university 
of Washington ¯ 
Madison Park 
Washington 
Pa rk 
aÅit61 Hill 
Madison valley 
Yarrm•' Point -z- 
Clyde Hill' 
Medina- 
•First ill 
Seattle 
Internationa 
DistricOChinatown 
iMadrona 
Leschi 
Homer 
Bridge 
Bellevue 
Beaux*rt5 
Beacon kill 
Rainier Valley 
Mercer Island 
Mercer 
Island 
Geoigetown 
62 

We will use the following models:

1. **K-Nearest Neighbor (KNN)** 
   1. KNN will predict the severity code of an outcome by finding the most like data point within k distance.
2. **Decision Tree** 
   1. A decision tree model will give the layout of all possible outcomes, so the model predicts all the different consequences of a decision. The decision tree observes all possible outcomes of different weather conditions.
3. **Logistic Regression** 
   1. As the dataset only has two severity code outcomes, the model will only predict one of those two classes. This makes the data binary, which is perfect to use with logistic regression.

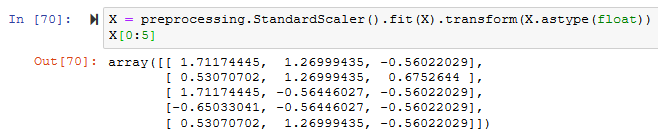
**Initialization**

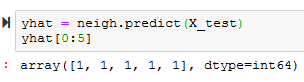
**Normalize the dataset**

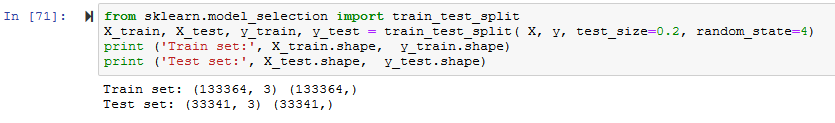


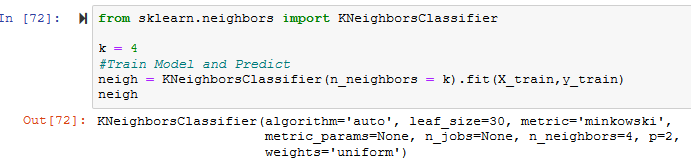
**Train/Test Split**

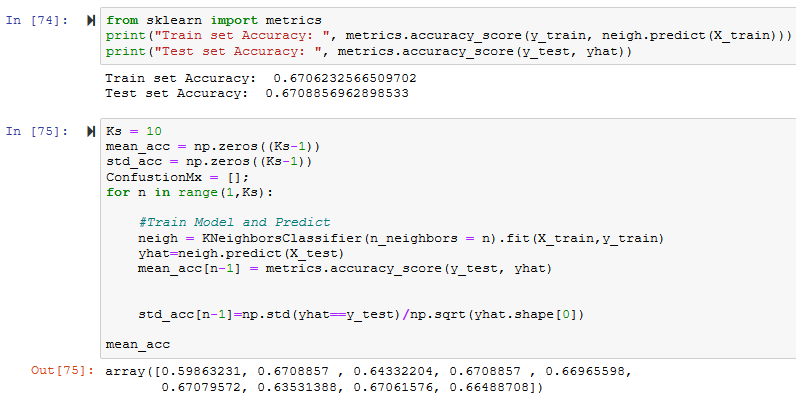
We will use 30% of our data for testing and 70% for training.

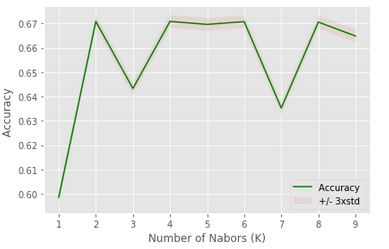




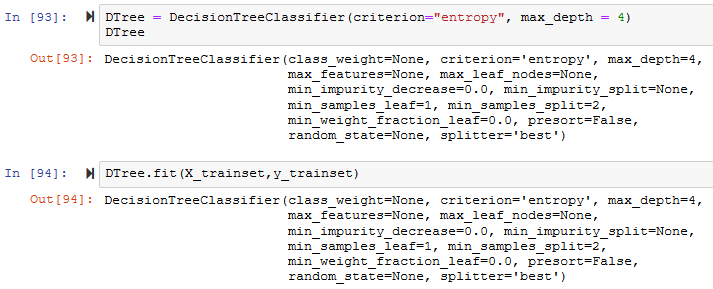




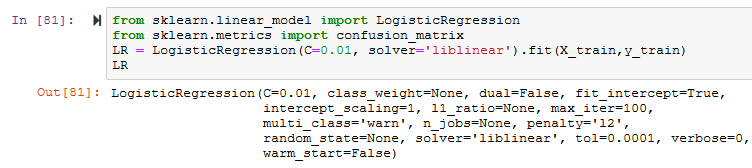


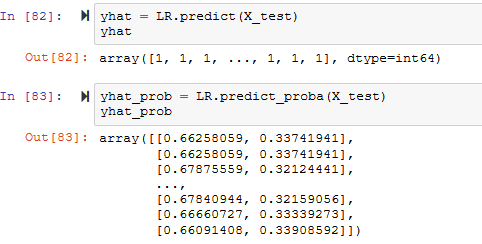


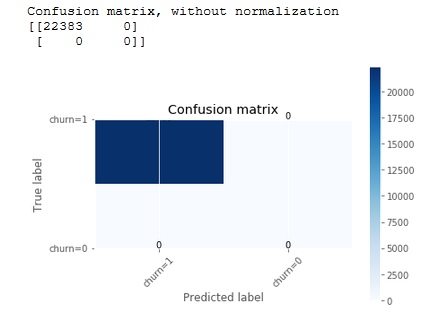








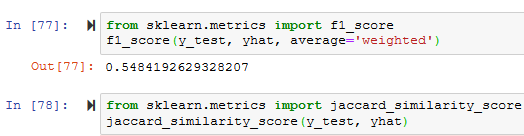




1. **Results section**

The accuracy of the three models









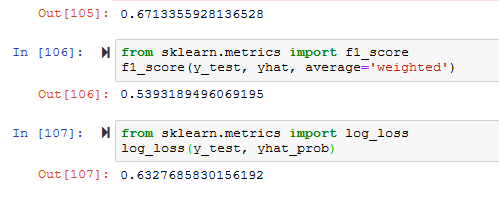












1. **Discussion section**

After analyzing and cleaning the data, it is fed through three ML models namely K-Nearest Neighbor also known as KNN, Decision Tree and Logistic Regression. The evaluation metrics for these three models that test the accuracy of the models are the Jaccard index, f-1 score and logloss for logistic regression.

1. **Conclusion section**

What is the best model? The quality of the model shouldn’t be only measured by accuracy rate, but the simplicity, easily understandable by decision maker and convenience to implement do matter the most

Logistic regression showed good performance than the other two models and can be the best model to implement for the reduction of fatal/ injury accident. There is no perfect solution, only a solution that is good enough for the intended purpose. This purpose can –and in many cases should– grow in complexity and sophistication as the results prove more and more useful and provide a feedback loop on how to improve themselves.

Thanks for reading!