IBM Data Science Professional Certificate Capstone Project

Predicting the Severity of Car Accident

by Bigboy Mutichakwa

September 2020

1. Introduction

1.1 Background

According to the World Health Organization (WHO), approximately 1.35 million people die from road traffic crashes all over the whole world every year. This being a cause for concern globally, the United Nations' Sustainable Goals' (SDGs) targets relating to road safety are to halve global fatalities resulting from road accidents by year 2020 (target 3.6) and *to provide access to safe and sustainable transport for all by 2030* (target 11.2).

The impact of road accidents is very huge ranging from human suffering caused from injuries and deaths to economic burdens imposed on the society to treat the injured and to take care of the deceased's dependents without mentioning the loss of productivity from those disabled and deceased.

1.2 Problem

The International Federation of Red Cross and Red Crescent Societies in their handbook of 2007 on road safety guidelines highlighted that road accidents are not only largely preventable but also largely predictable. Wouldn't it be a good thing if we could use the data that we have collected from past accidents to predict the severity of accidents using supervised machine learning algorithms for the emergency response

teams to mobilize the required amount of resources for a particular accident occurrence?

Data about roads, users, vehicle conditions might be relevant in predicting the occurrence of an accident and its severity. According to the Haddon Matrix, one of the most popular models that is used in the field of injury prevention (Wikipedia), environmental (road design, speed limits and pedestrian facilities), vehicle and equipment (speed management, lighting, braking, road worthiness) and human factors (police enforcement, attitudes, impairment) are identified as key factors to consider before an accident occurs (pre-crash phase). Some of these factors are also recognized by the United Nations in its efforts to curb road traffic accidents globally. To achieve SDG target 3.6, the United Nations identified key areas that need attention some of which include planning, designing and maintaining a safe road infrastructure and promoting high safety standards for new cars.

1.3 Interested stakeholders

Accident severity prediction assist government authorities and health officials to activate the right emergence response procedures just at the time of an accident occurring so that the right resources are mobilized and dispatched on time. I am also persuaded to believe that such information would certainly come in handy for road traffic users to either cancel certain trips or change routes where possible. Road administration authorities could also use the same information to predict and warn road users or advise on other uncongested routes basing on the predicted severity just at the time of the accident occurring.

2. Data

Our data from Seattle's transport department was downloaded from Seattle's website. It contains features on collisions in the city of Seattle, a seaport city on the West Coast of the United States of America from 2004 to 27 August 2020. This data is available here.

The dataset we are working with contains 221266 records and 40 columns.
SEVERITYCODE is going to be our label column.

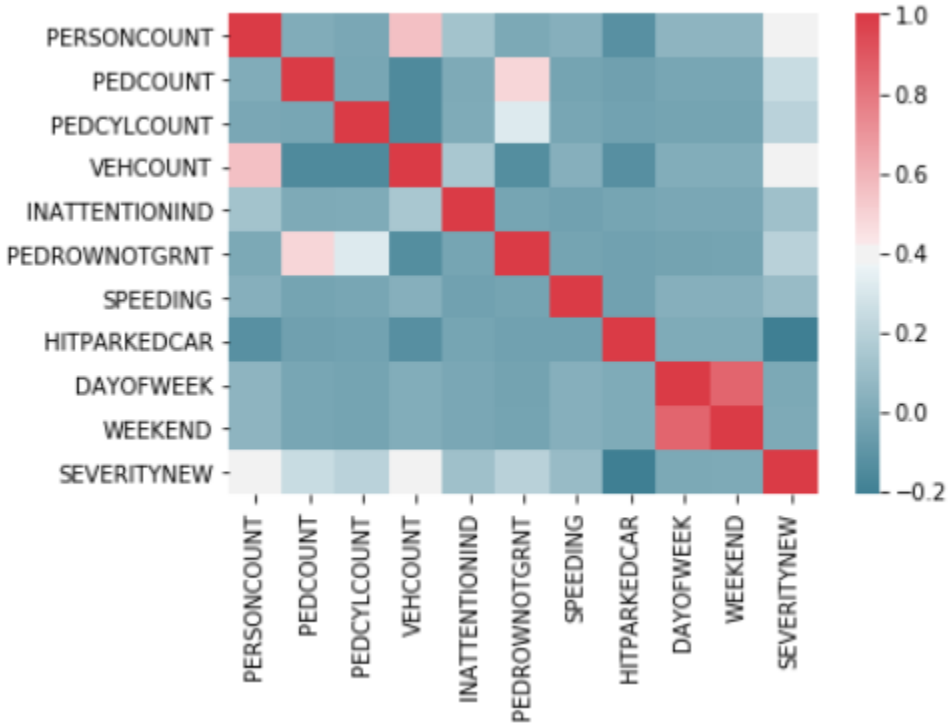
2.1 Removing features which are not required

The following features were seen to be not adding value when it comes to developing our predictive model and as such are removed.

- **OBJECTID, X, Y** : the ESRI unique identifier and location coordinates
- **INCKEY, REPORTNO, COLDETKEY, INTKEY** : a unique key for the incident, report number, secondary key for the incident and key referencing collision intersection respectively
- **STATUS** : it is not clear what information is stored in this column since it also doesn't appear in the metadata file of the dataset
- **LOCATION** : description of the location of the collision
- **EXCEPTRSNCODE,EXCEPTRSNDESC** :it is not clear what information is stored in these columns since no descriptions appear in the metadata file of the dataset
- **SEVERITYDESC** :this is simply a description of collision severity code (label) and will not be useful to keep it.
- **INJURIES, SERIOUSINJURIES, FATALITIES** : this information will not be available at the time of the collision at which point we want to predict how severe the collision is.
- **JUNCTIONTYPE** : this column is derived from the ADDRTYPE (collison address type). We will remove this column and retain the ADDRTYPE column.
- **SDOT_COLCODE, SDOT_COLDESC**: the code and description given to collision will only be available after the collision and hence these columns will not help us to predict the collision severity.
- **COLLISIONTYPE** : collision type is correlated with the collision severity that we want to predict.
- **SDOTCOLNUM** : collision number will not be able to assist in predicting the severity level of an accident.

- **ST_COLCODE, ST_COLDESC** : state collision code and description which will be available after collision has been attended to by state will not be useful to us to create a predictive model.
- **SEGLANEKEY, CROSSWALKKEY** : lane segment and crosswalk keys are not relevant for our predictions.

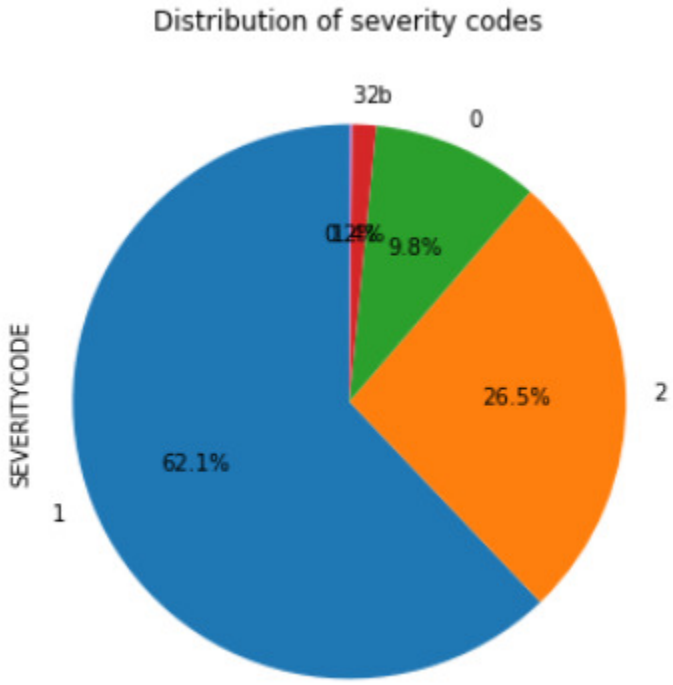
2.2 Correlation heatmap



Although the above heatmap drawn by the seaborn Heatmap function is not a perfect tool, we can have some better understanding at a glance that the vehicle count (VEHCOUNT) and the total number of people involved in a collision (PERSONCOUNT) has some positive correlation with the severity of the collision. A negative correlation exists between the severity of a collision and a collision which involves hitting a parked car.

2.3 Distribution of severity codes

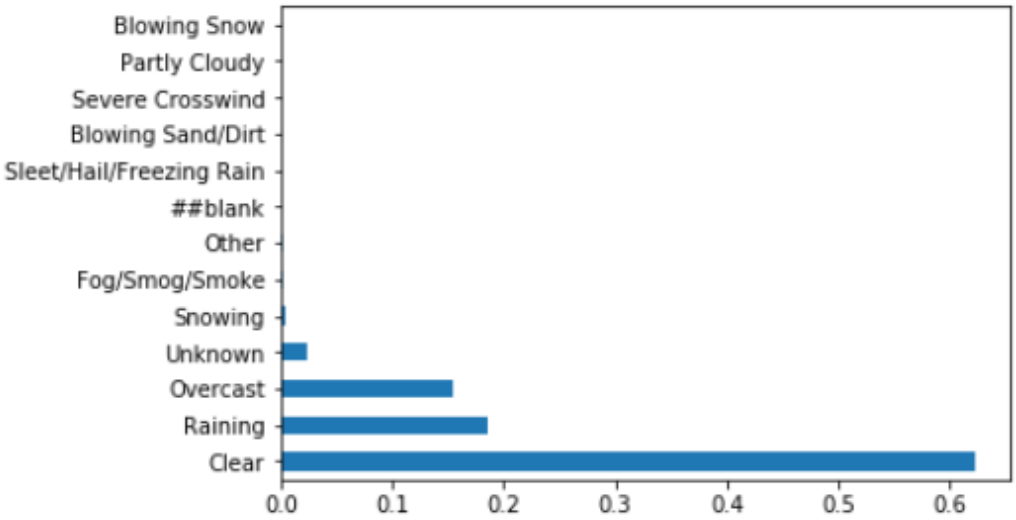
62.1% of all the accident records are for severity level 1 (property damage) and the smallest percentage is 0.2% which is for severity level 3 (fatal).




```
1     62.1%
2     26.5%
0      9.8%
2b     1.4%
3      0.2%
Name: SEVERITYCODE, dtype: object
```

2.4 Null values and unknowns

Every column was checked for nulls and guided by the proportion of the null values within each particular column, the nulls were removed or treated. For example, null values in the WEATHER column constituted less than 5% of the rows and as such a decision to proceed without these rows was made.



2.5 Normalizing Data

The feature columns were normalized using StandardScaler function such that the distribution of the data for every column will have a mean of 0 and a standard deviation of 1.

Normalizing Data

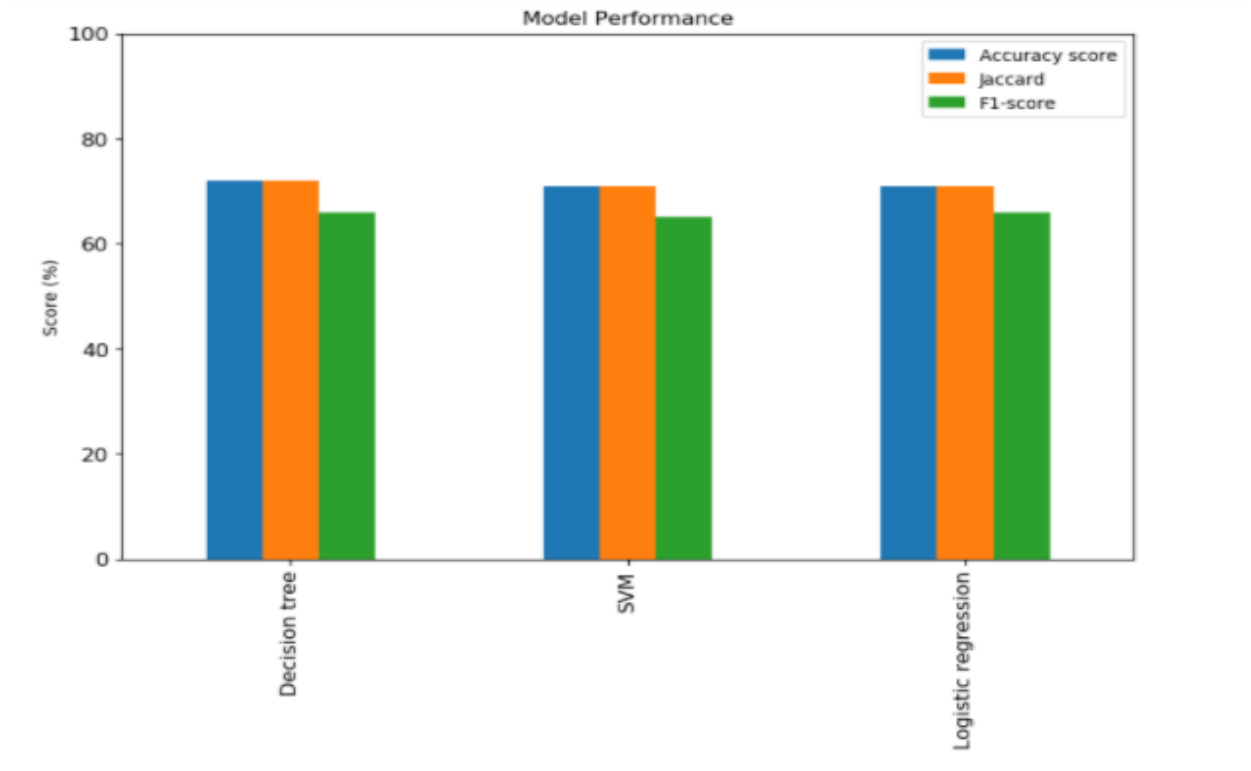
```
In [106]: X= preprocessing.StandardScaler().fit(X).transform(X)
          X[0:5]

Out[106]: array([[-3.54736071e-01,  4.29135074e+00, -1.84253374e-01,
                  -1.62973016e+00, -4.45404132e-01,  5.83242896e+00,
                  -2.42079880e-01, -1.72284236e-01, -4.90832383e-01,
                  -8.50917781e-01, -5.90298279e-02, -1.31130881e+00,
                   1.32115749e+00,  2.39500396e-01, -2.39487044e-01,
                  -2.39179016e-03, -2.58797545e-02, -1.56859015e-02,
                  -2.39179016e-03,  7.51535504e-01, -5.65382615e-02,
                  -3.88174191e-02, -4.33525402e-01, -7.17553467e-03
```

3. Predictive modelling

Three supervised machine learning models were selected to create predictive models using the normalized data. The best performer was then recommended.

- **Decision tree:** one of this model's biggest strength is that it is not easily affected by outliers and it learns non-linear relationships well. However, one of its downsides is that it tends to overfit on training data.
- **Support Vector Machine (SVM):** SVM model is opted due to the model's ability to handle data with large number of feature attributes and has a strong capability to generalise.
- **Logistic regression**: this model is simple to implement and prevents over-fitting although over-fitting is highly possible in high dimensional planes.

3.1 Performance report



Algorithm	Accuracy score	Jaccard	F1-score	LogLoss
Decision Tree	0.72	0.72	0.66	NA
SVM	0.71	0.71	0.65	NA
LogisticRegression	0.71	0.71	0.66	0.63

From the evaluation metrics above, we can see that the Decision Tree model has the highest accuracy score although the differences (72% vs 71% for other models) indicate that all models have generally the same level of performance. The Logistic regression model has a generally high (63%) log loss which indicates that the predicted probability is 63% far from actual labels on average. Efforts should be made to reduce the log loss and increasing the accuracy score.

4. Conclusion

Previous studies have indicated that location type is a good feature to predict accident severity. There is therefore room to improve the models by incorporating the geolocation data columns into building the models. These columns can be used by zoning the geolocation columns and then hot encode the categorical zone column.

References

- Blogspot. (2019) Advantages and Disadvantages of Logistic Regression in Machine Learning. [Online] Available from http://theprofessionalspoint.blogspot.com/2019/03/advantages-and-disadvantages-of.html
- EasyAI. (2019) Support Vector Machine. [Online] Available from https://easyai.tech/en/ai-definition/svm/
- Plos One Journals. (2019) A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. [Online] Available from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0214966#
- Medium. (2019) Using Latitude and Longitude data in my machine learning problem. [Online] Available from https://medium.com/@khadijamahanga/using-latitude-and-longitude-data-in-my-machine-learning-problem-541e2651e08c/