IBM Data Science Professional Certificate Capstone Project on Coursera

Development of an artificial intelligence model to predict the severity of car crashes.  
  
Daniel Asante Otcherea  
a Department of Petroleum Engineering, Universiti Teknologi PETRONAS, 32610, Seri Iskandar, Perak Darul Ridzuan, Malaysia

Executive Summary

Road accidents are important topics of our current civilisations, accountable for millions of deaths and injuries every year in the world. In this research, factors impacting the severity of road accidents are analysed. Two machine learning algorithms were used to classify road accidents. It was found that road accidents occurs mostly due to poor raod, lighting and weather conditions. Also, road users also contribute to road accidents mostly at intersections. The KNN model recorded an F1 score of 71 % and an accuracy score of 75% making it the most suitable classification algorithm for this research.

# **Introduction**

Road safety is a major concern for governments as well as major automobile corporations because there are over 38,000 people die every year in crashes on U.S. roadways out of which about 9 people die each day. There is an additional 4.4 million people who get seriously injured to require medical attention. Tech companies are also investing huge amounts of money in autonomous cars. Therefore, it is advantageous to look at car safety from a data perspective and determine what factors contribute to road accidents. We also look at the severity of accidents and build a model that predicts the severity of a road accident.

There are a number of factors that contribute to road accidents resulting in deaths and severity of injuries. These factors are:

1. Poor road infrastructure and management.
2. Non-road worthy vehicles.
3. Unenforced or non-existent traffic laws.
4. Unsafe road user behaviours.
5. Inadequate post-crash care.

By understanding each of these factors and through planning, effective management and evidence-based interventions, road crashes can be predicted and prevented. Having access to accurate and updated information about the current road situation enables drivers, pedestrians and passengers to make informed road safety decisions. With the death toll of road accidents going up, governments as well as major automobile corporations are majorly concerned. For road users, be it cars or any vehicle that plies the roads, it is highly important to look at the safety of the vehicle from a data point of view and determine what factors contribute to road accidents. This project also assesses at the severity of accidents and build a model that predicts the severity of a road accident that will help commuters to decide in choosing an alternate route.

## **Criteria for selecting best model**

In model classification, the difference between the actual data points and the best fit line produced by the algorithm is the model’s error. In this project the accuracy and similarity score will be used to evaluate the model. In classification, accuracy classification score is a function that computes subset accuracy. This function is equal to the jaccard similarity score function. Essentially, it calculates how closely the actual labels and predicted labels are matched in the test set.

**Materials and Methodology**

A typical model development workflow used in this study is as follows;

1. Define the aims and objectives for model.
2. Analyse current system.
3. Data analysis and pre-processing.
4. Develop a model for training and testing.
5. Validate model against data not seen by the model.
6. Deploy model.

Figure 1 Sequence of a typical machine learning workflow

Data provided by Coursera for this capstone was used for this project. The data has lots of variables but only a few were selected for the model. The input variables used were dependent on their correlation to the target label which is the severity code. The severity code is classified as accident type: Property damage with severity code of 1 and Injury with severity code of 2

The input variables used are collision type, vehicle count, pedestrian count, weather, road condition, light condition, speeding and junction type. Table 1 gives a brief recap of the statistical description of some of the input data (data type = integer) used to train the models.

Table 1 Descriptive statistics of some input variables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **SEVERITYCODE** | **PERSONCOUNT** | **PEDCOUNT** | **VEHCOUNT** |
| Count | 194673 | 194673 | 194673 | 194673 |
| Mean | 1.298901 | 2.444427 | 0.037139 | 1.92078 |
| Std | 0.457778 | 1.345929 | 0.19815 | 0.631047 |
| Min | 1 | 0 | 0 | 0 |
| 25% | 1 | 2 | 0 | 2 |
| 50% | 1 | 2 | 0 | 2 |
| 75% | 2 | 3 | 0 | 2 |
| Max | 2 | 81 | 6 | 12 |

## **Data analysis and preprocessing**

A series of data cleaning techniques were employed to change data types and identify missing data. Plots and correlations from the seaborn library were used to analyse the data. From the plots and charts below, features that has a high impact on the severity code were identified.

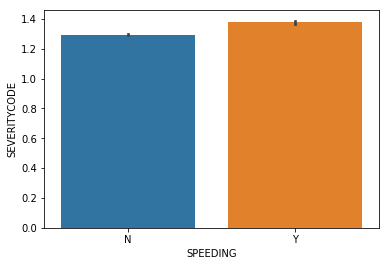


Figure 2 Impact of speeding on target label

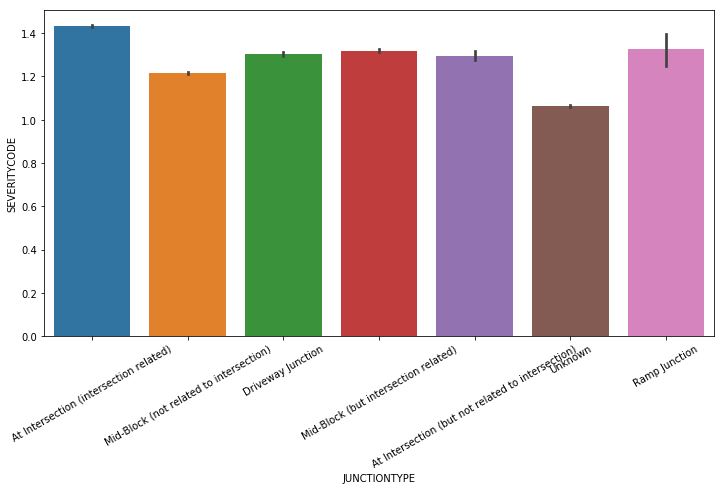


Figure 3 Impact of the type of junction on target label

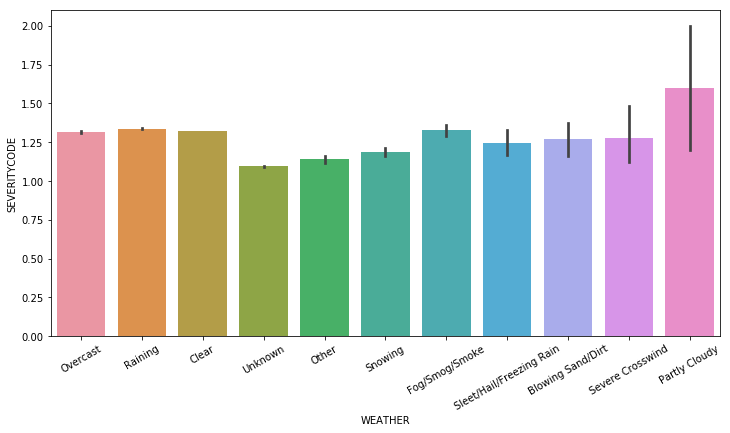


Figure 4 Impact of weather type on target label

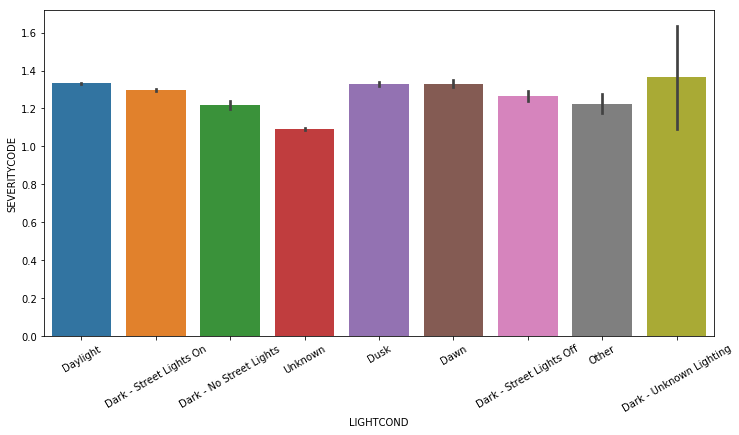


Figure 5 Impact of Lighting condition on target label

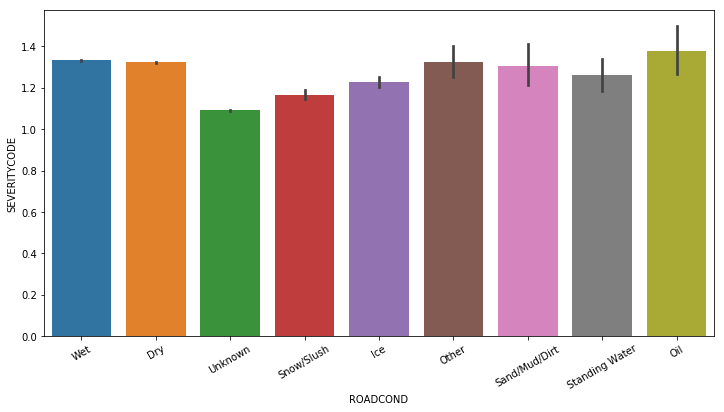


Figure 6 Impact of Road condition on target label

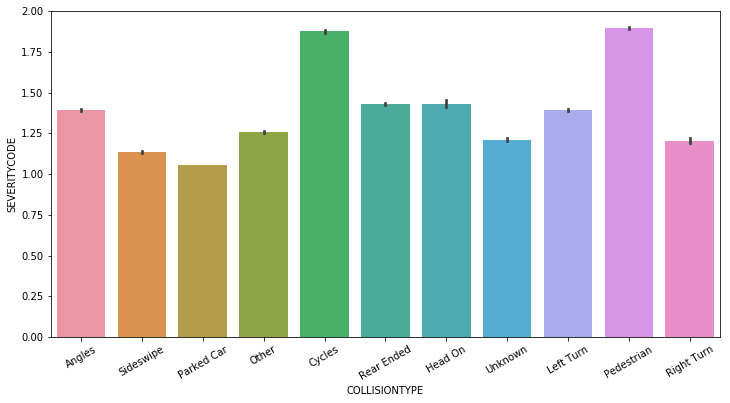


Figure 7 Impact of collision type on target label

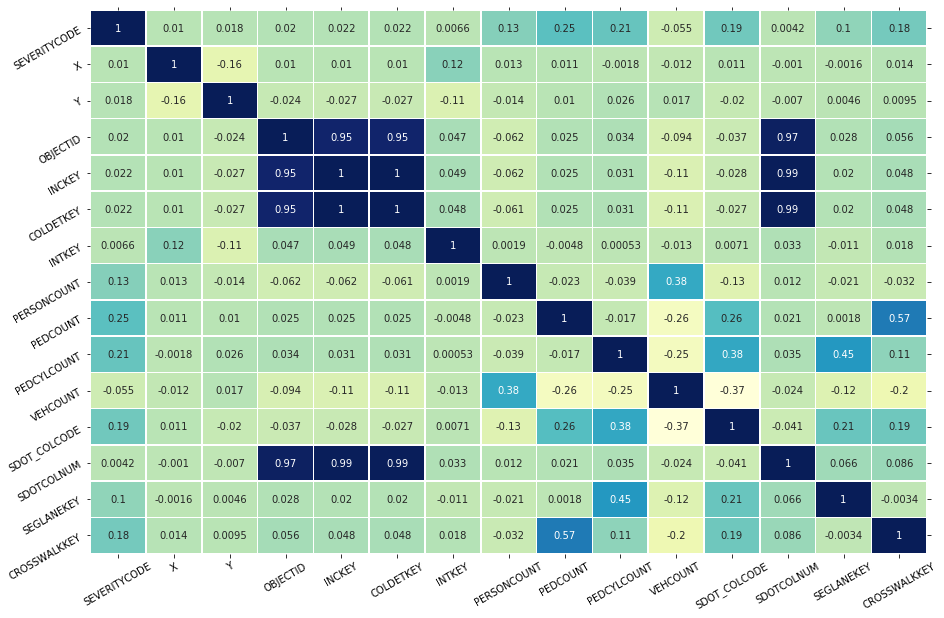


Figure 8 Pearson correlation between each parameter

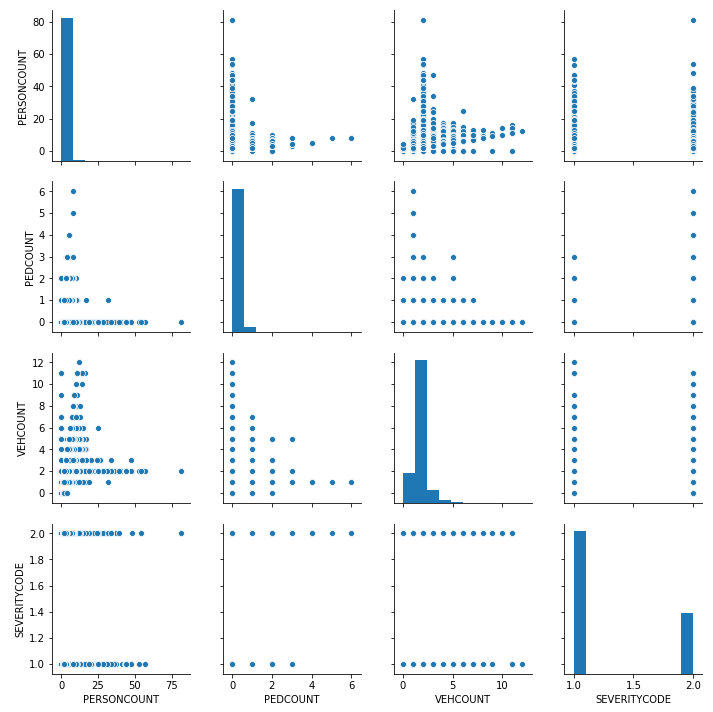


Figure 9 Matrix of crossplots to visualize the correlation between the parameters and target label.

## **Machine learning models used**

The machine learning models used in this classification are the support vector machine (SVM) and K nearest neighbour (KNN). SVM is a supervised algorithm that classifies cases by finding a separator. It works by matching data to a high dimensional feature space so that data points can be categorised even when the data are not otherwise linearly separable. A separator is then estimated for the data. The data is transformed in a way that a separator could be drawn as a hyperplane. The main advantages of the SVM are that it is accurate in high-dimensional spaces and uses a subset of training points in the decision function called support vectors making it memory efficient. The main disadvantages of this algorithm is that it is prone to overfitting if number of features are greater than number samples, it does not provide probability estimations which are desirable in most classification problems. The data for SVM was split into a ratio of 70:30 for training and testing respectively. KNN is a classification algorithm that takes labelled points and uses them to learn how to label other points. This algorithm classifies cases based on their similarities to other cases. In KNN, data points that are near each other are said to be neighbours. KNN is based on the paradigm that similar cases with the same class labels are near each other thus the distance between two cases is a measure of their dissimilarity. One way to calculate the distance or dissimilarity of two data points is by using the Euclidean distance. The grid search cross validation function was used to split the data using 5 fold cross validation with the best possible number of neighbours selected.

With the SVM different kernel functions were used and compared to select the most accurate. The best model was then compared to KNN to select the final model.

# **Results**

The SVM-RBF recorded an average F1 score of 0.69 and similarity score of 0.75 whiles the SVM-Sigmoid kernel function recorded average F1 score of 0.64 and similarity score of 0.64. The SVM-linear also recorded an average F1 score of 0.64 and similarity score of 0.73. KNN model gave average F1 score of 0.71 and similarity score of 0.75.

Table 2 Summary results of machine learning models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | SVM-RBF | SVM-Sigmoid | SVM-Linear | KNN |
| F1 Score | 0.69 | 0.64 | 0.64 | 0.71 |
| Accuracy | 0.75 | 0.64 | 0.73 | 0.75 |

# **Discussion**

From the data analysis stage, it was found that speeding had slight impact on the severity code which is mainly due to missing data. For the junction type, the severity code was highly impacted by intersections (intersection related). The severity code also is affected in partly cloudy weather as well as dark lighting conditions. For road conditions, oil had a high effect on the severity codes whiles the collision type was mostly pedestrians involved in the accidents. The above analysis points to the fact that under poor weather and road conditions, accidents have a high likelihood of occurrence whiles pedestrians are involved in most accidents which occurs at intersection mainly because drivers and pedestrians tend not pay attention to the road when entering an intersection or joining the main road.

From the machine learning model, was ranked as the best classification algorithm based on the criteria assessment used. The 5-fold cross validation technique may have contributed to this. This does not mean it is better than the SVM but that SVM tends to perform poorly when the data used is large (over 1000 rows).

# **Summary and Conclusion**

The project helps identify factors that contribute significantly to road accidents using a data-driven decision-making process. It can be concluded that poor road, weather and lighting conditions lead to most accidents whiles road users at intersections also tend to contribute to road accidents. Speeding may have had an inconclusive determinant on road accidents but thi is due to missing data in the dataset used. Data such as alcohol tests was not part of the dataset and would have been interesting to see its effect on road accidents. Based on this dataset the KNN established itself as a better machine learning model in road accident classification.

Acknowledgement

The author expresses his sincere appreciation to IBM and Coursera for making this course available and for financially aiding the whole professional certificate course.