**Abstract**

Having roads free from fatalities and serious injuries by 2028–Vision Zero–is what the City of Alexandria is working towards. This project is about using historical traffic data for the City of Alexandria to determine if Vision Zero will be achieved by 2028. I follow the CRISP-DM framework to guide my analysis, model evaluations and selection, and predictions. I conclude that, given historical data, the City of Alexandria will not achieve zero fatalities and serious injuries from traffic accidents by 2028. I recommend that certain important features should be investigated further to understand areas for improvement if the City of Alexandria intends on staying on course to achieve Vision Zero by 2028.

**Introduction**

Fatalities and serious injuries due to traffic accidents can move decision makers and stakeholders to take drastic steps to change things for the better. One such drastic step is the City of Alexandria's aims to have zero traffic fatalities and serious injuries by 2028. This date may sound far-off or too ambitious. Cities around the U.S. have caught on to Vision Zero, an initiative started in Sweden in 1997 to have no deaths and serious injuries due to car accidents. Is this achievable?

To understand whether Vision Zero is achievable we need to look at what historical traffic data are telling us. This is the scope of this project: to use publicly available historical traffic data from the Virginia Department of Transportation to predict whether the City of Alexandria will achieve Vision Zero by 2028.

**Problem Definition and Formulation**

I followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework to structure my data mining workflow. CRISP-DM includes project understanding, data understanding, data preparation, modelling, evaluation, and deployment.

<u>Project understanding</u>

Will the City of Alexandria achieve zero traffic fatalities and serious injuries by 2028? The answer to this question will be a plot that visualises historical traffic fatalities and serious injuries for 2015-2022 along with predictions for 2023-2028. The expected benefit of this project could help decision-makers and stakeholders scrutinise traffic features to understand

which parameters are helping or hindering the City of Alexandria achieving Vision Zero by 2028.

Data understanding

I used crash data from the Virginia Department of Transportation's Open Data Portal. The data are updated regularly and span crash incidents since 2015. The data set had 882,957 observations and 40 variables at the time I downloaded it. Some attributes include:

- Crash date;
- Crash military time;
- Number of drivers;
- Driver action;
- Driver(s), passenger(s), pedestrian(s) ages;
- Driver(s), passenger(s), pedestrian(s) injury type;
- Crash severity;
- Collision type;
- Vehicle body type;
- Light condition;
- Weather condition, etc.

The granularity was daily, but not evenly spaced to warrant time-series analysis.

Data preparation

This project is about the City of Alexandria. So, I filtered the imported data set for Alexandria and ended up with a data set with 10429 observations and 40 variables. After that, I filtered by crash severity being either a severe injury or fatality. This reduced the data set to 282 observations and 40 variables–the final data set that I labelled, 'alex'.

I enhanced the data quality by cleaning and formatting variables according to what would be a sensible data type for each variable. Missing values were mostly string values that were not recorded. I filled in these missing string values with "Not Recorded"–in line with how the original data set recorded some missing or unknown values.

The data had multiple instances where cells contained different data types. Numeric data, for example, were wrapped inside character data (quotation marks, commas, semi-colons) in the same cell. I removed leading, middle, and trailing white spaces.

I assessed skewness and kurtosis quantitatively (with skewness and kurtosis functions) and visually (with boxplots and histograms). I performed transformations to standardise skewed variables (as well as variables with above normal excess kurtosis).

I used decision trees to narrow down the following important variables:

- Route name;
- Collision type;
- Driver action;
- Total drivers;
- Vehicle type;
- Light condition;
- Road surface condition;
- Driver gender.

Once I prepared the data and narrowed the variables down, I copied the data set and made 'serious injuries' (*serious*) and 'fatalities' (*fatal*) the only target variables in each respective data set. That is, 'serious injuries' was the only target variable in one data set without 'fatalities', and vice versa. This way I can tailor models based on two separate data sets.

I summed the target values by month, therefore making the granularity monthly, because the daily granularity of the original data were inconsistent. Since the target variables capture and represent the outcome of the eight important variables (selected by decision trees), I narrowed the *serious* and *fatal* data sets down to only 'Crash date' and 'Serious injuries' columns, and 'Crash date' and 'Fatalities' columns respectively.

Modelling

I split the *serious* and *fatal* data sets into a 70% training set and 30% testing set. I first fit a linear model to determine the conditions for linear regression (constant variance, errors are approximately normally distributed). The data sets failed to meet these conditions, so I

decided on models that are suitable for non-linear regression. I fitted the following models to the training sets:

- Support Vector Machines (SVM)
- Random Forests
- Multivariate Adaptive Regression Splines (MARS)
- Auto-Regressive Integrated Moving Average (ARIMA)
- Neural Network (single layer)
- Deep Neural Network (three hidden layers)

Evaluation

I predicted on the training and test sets for *serious* and *fatal* and used the root mean squared error (RMSE) to compare the models. I also calculated the difference between the test RMSE and train RMSE. This helped me understand the extent of overfitting. If the difference between the training and testing RMSEs was large, then it suggests possible overfitting.

I tabulated all the models for *serious* and *fatal* and sorted the tables first by the lowest `Test.RMSE` then by the lowest `Train.Test.Difference`. For *serious* I chose the second SVM; for *fatal* I chose the second neural network (deep neural network).

*serious*

| Model | Train.RMSE | Test.RMSE | Train.Test.Difference |
|---|---|---|---|
| serious.SVM2 | 2.186470 | 2.205921 | 0.01945047 |
| serious.RF | 1.369182 | 2.247285 | 0.87810360 |
| serious.MARS | 2.338538 | 2.374308 | 0.03577018 |
| serious.ARIMA | 2.770964 | 2.387267 | -0.38369754 |
| serious.SVM1 | 2.366170 | 2.426237 | 0.06006623 |
| serious.NN2 | 2.440662 | 2.520752 | 0.08008950 |
| serious.NN1 | 3.498518 | 3.642651 | 0.14413362 |

*fatal*

| Model | Train.RMSE | Test.RMSE | Train.Test.Difference |
|---|---|---|---|
| fatal.NN2 | 0.7916862 | 0.5683812 | -0.22330497 |
| fatal.MARS | 0.8017837 | 0.5838742 | -0.21790952 |
| fatal.NN1 | 0.8017837 | 0.5838769 | -0.21790682 |
| fatal.SVM1 | 0.9047092 | 0.6342331 | -0.27047602 |
| fatal.SVM2 | 0.8436630 | 0.6695157 | -0.17414723 |
| fatal.RF | 0.5353803 | 0.8114454 | 0.27606510 |
| fatal.ARIMA | 1.0273905 | 0.9655005 | -0.06188995 |

**Solutions and Discussion**

I created an empty data frame with future dates starting at `2022-01-01` up to `2029-01-01`.

```
# Create start and end dates
start.date <- ymd('2022/01/01')
end.date <- ymd('2029/01/01')

# Predict serious injuries for 2022-2029
serious.2028 <- data.frame(CrashYear = seq(start.date,
end.date,'months'))
```

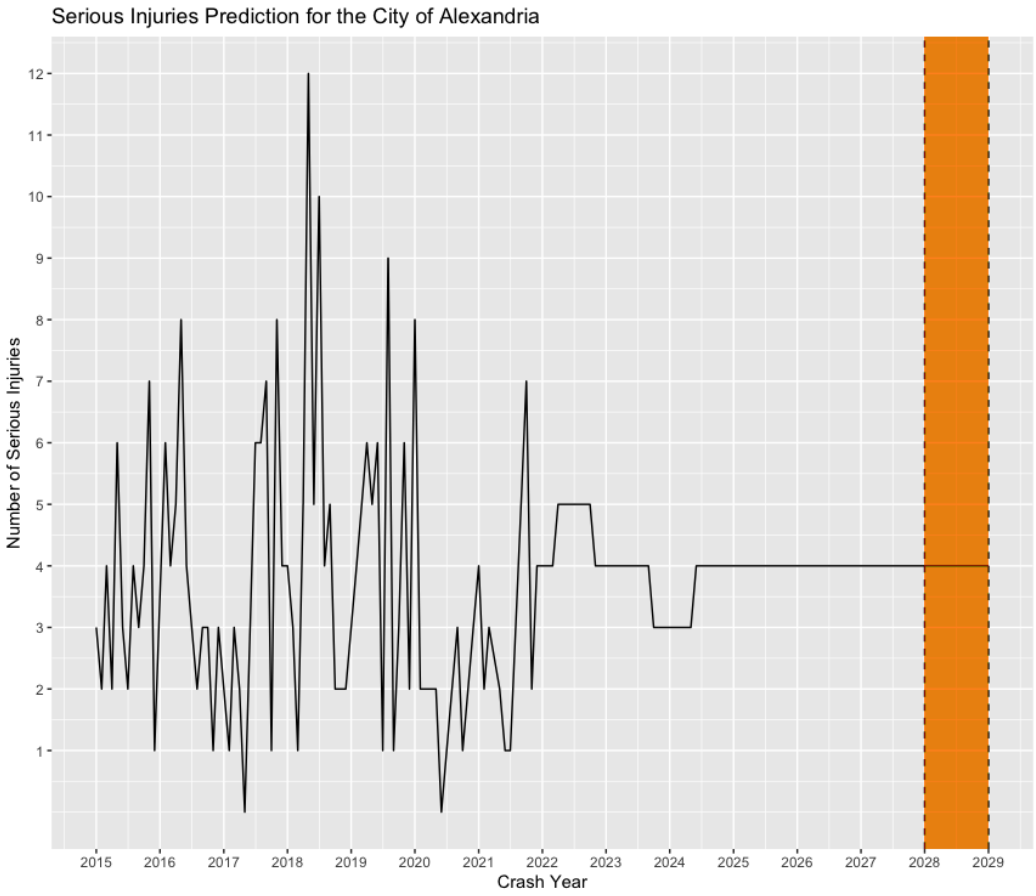
I then used my respective models for serious injuries and fatalities and predicted over these future dates. I rounded decimals that were ≥ `0.5` up:

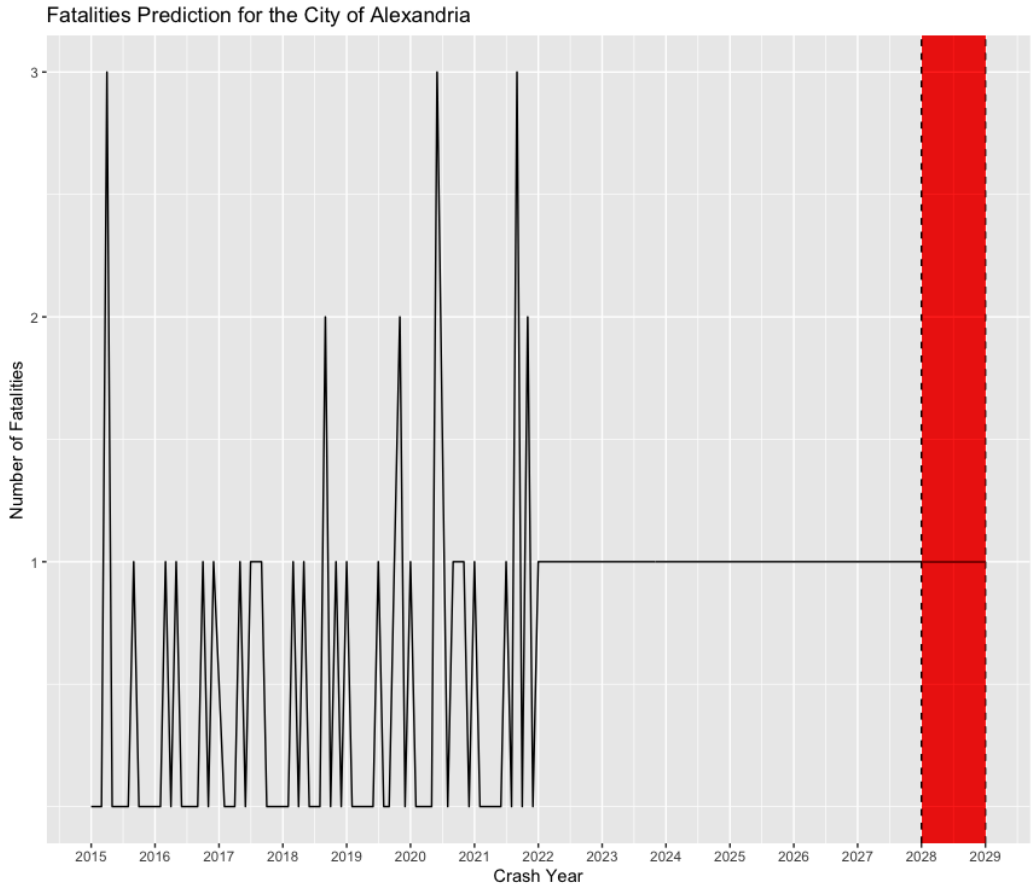
```
serious.predictions <- predict(serious.model,
serious.2028)
serious.predictions <- floor(0.5 + serious.predictions)
# round up numbers with decimals >= 0.5
serious.predictions # sanity check

# Predict fatalities for 2022-2029
fatal.2028 <- data.frame(CrashYear = seq(start.date,
end.date,'months'))
fatal.2028

fatal.predictions <- predict(fatal.model, fatal.2028)
fatal.predictions <- floor(0.5 + fatal.predictions)
# round up numbers with decimals >= 0.5
fatal.predictions # sanity check
```

I added these predictions back to their respective data sets–*serious* and *fatal*–and plotted the results:


Serious Injuries Prediction for the City of Alexandria
Number of Serious Injuries
12
11
10
9
8
7
6
5
4
3
2
1
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
Crash Year


Fatalities Prediction for the City of Alexandria
Number of Fatalities
3
2
1
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
Crash Year

The vertical shaded bands (orange for serious injuries; red for fatalities) highlight the year–2028–that the City of Alexandria aims to achieve Vision Zero. The interpretation of these plots suggests that there will be approximately four serious injuries per month and approximately one fatality per month in 2028. It is evident that the City of Alexandria will not achieve Vision Zero by 2028.

Recommendations to decision makers and stakeholders include further investigation of the eight important variables (determined by the decision tree) to assess areas for improvement. Some recommendations are beyond the scope of this project. However, recommendations include investigating the correlation between important variables such as specific routes, road surface conditions, and initiatives to change driver behaviour.

**Conclusions**

The data set by the Virginia Department of Transportation needed to be prepared to select the best-performing models–support vector machines and deep neural networks–to predict and provide clarity for decision makers and stakeholders of the City of Alexandria towards achieving Vision Zero.

**References**

Can Better Data Make Zero Traffic Deaths a Reality? [accessed 2022a May 03]. https://datasmart.ash.harvard.edu/news/article/can-better-data-make-zero-traffic-deaths-a-reality-1138.

Virginia Roads. [accessed 2022b April 12]. https://www.virginiaroads.org/.

Vision Zero. City of Alexandria, VA. [accessed 2022c April 07]. https://www.alexandriava.gov/VisionZero.

Waze. 2020. Traffic Lab: Predicting a World Without Car Crashes. Waze. [accessed 2022 May 03]. https://medium.com/waze/traffic-lab-predicting-a-world-without-car-crashes-86e70126d6aa.