**Capstone Project**

**Document Skeleton**

# Process overview

The following diagram shows the overall end-to-end process for defining, designing and delivering the Capstone project.



Note: The following are the candidate sections of the document. They are presented here for guidance. Questions in each section could be used as possible aspects to cover. Some questions may not be applied to each project. On the other hand, additional information may be needed.

# Problem statement

* What is the problem or the opportunity that the project is investigating?
  + Looking at factors that influence the road toll and whether or not we can predict targeted areas to reduce the road toll
* Why is this problem valuable to address?
  + Large cost to economy through property damage, healthcare costs, loss of potential earnings and emotional cost involved in the loss of life.
* What is the current state (e.g. unsatisfied customers, lost revenue)?
  + Lost potential future earnings affecting GDP, profitability of transport industry; grieving families due to lost loved ones
  + Road toll is decreasing year on year – perhaps it can be improved at a faster rate?
* What is the desired state?
  + Reduction of road fatalities incrementally and eventually down to zero (very long term)
* Has this problem been addressed by other research projects? What were the outcomes?
  + One study using artificial neural networks was used to predict serious injury road accidents in Malaysia with a mean absolute error of 3%.

# Industry/ domain

* What is the industry/ domain?
  + Government - State/Federal government; Transport and Freight Industry
* What is the current state of this industry? (e.g. challenges from startups)
  + Industry – Challenges from opposition/media/public pressure – funding and implementing measures to reduce the road toll; loss of productivity challenges due to road traffic accidents
* What is the overall industry value-chain?
  + "The Transport and Logistics sector in Australia has an estimated annual revenue of $102.87 billion, with an operating profit of $10.14 billion in 2018.” (AAA 2017) Some loss of employee for the transport industry would reduce this value. The loss of each person due to road fatality represents a loss of future earnings of 4.34m – affecting the earning potential of the nation.
* What are the key concepts in the industry?
  + Protecting life in the public using public health, policing and infrastructure measures. Protecting life in the transport industry due to safe work practices and awareness.
* Is the project relevant to other industries?
  + Trucking industry – loss of staff will affect productivity – might help identify the greatest risk profile for driving

# Stakeholders

* Who are the stakeholders? (be as specific as possible)
  + Government, police, ambulance, fire fighters (first responders), healthcare system, public.
* Why do they care about this problem?
  + Emotional reasons, cost of healthcare, loss of productivity for road fatalities, injuries etc.
* What are the stakeholders’ expectations?
  + Identify risk profile for driving that could lead to loss of life

# Business question

* What is the main business question that needs to be answered?
  + Can we use a prediction for future fatalities to set a target for reduction in road toll?
* What is the business value of answering this question? (quantify value and make necessary assumptions)
  + Approximately $22b, made up of lost potential earnings
* What is the required accuracy? What are the implications of false positives or false negatives?
  + False negatives are the concern – false positives will lead to more caution when driving. False negatives will lead to loss of life around that time – perhaps not a fluctuation from the norm but will remain along control group.

# Data question

* What is the data question that needs to be answered?
  + Can we predict road fatalities in the future and can we produce a campaign to impact a reduction in the road toll?
* What is the data required to answer the question?
  + Details of the fatal accidents that have happened in Australia in the past.

# Data

* Where was the data sourced?
  + Data.gov.au
* What is the volume and attributes of the data?
  + 52,397 observations over 30 years. 22 features – time of incident, day of the week, month, age of victim, gender, speed limit, car involvement (single/multiple), bus/truck involvement, road user type, holiday periods (Easter/Christmas), road types
* How reliable is the data?
  + Very reliable – produced by government and constantly updated (monthly and quarterly for different measures)
* What is the quality of the raw data?
  + Quality is pretty good – some features have a lot of missing values, most of the data is there
* How was this data generated?
  + Based on reports of vehicle accidents causing fatalities
* Is this data available on an ongoing basis?
  + Yes – always available and free on data.gov.au

# Data science process

## Data analysis

* What data pipeline was to wrangle the raw data?
  + The pipeline involved acquisition of the data, cleaning to remove null values, I removed a couple of columns with a large proportion of null values, then some feature engineering to turn the months in a column of MM-YYYY for ease of use during the modelling stage. A lot of the visualisation is on the entirety of the cleaned dataset, which involved a lot of visualisation of value\_counts() rather than raw data/single observations/rows.
* What are the highlights of the Exploratory Data Analysis (EDA)?
  + Highlights are the different days of the week / month of the year and time of day that are most conducive to fatal accidents: the weekend; December and March; around 1400-1800 in the afternoon.
* Is the pipeline reusable? (for example, to process future data?)
  + No, I don’t believe so. Additional data will need to be manually cleaned to be usable in the future.
* What are the intermediary data structures used (if any)?
  + I changed notebooks several times after I had earlier issues loading a notebook that had become too large of a filesize

## Modelling

* What are the main features used?
  + Monthly count of fatalities over the timespan
* Did you find any interesting interactions between features?
  + None, as I was only using one feature to predict future fatalities
* Is there a subset of features that would get a significant portion of your final performance? Which features?
  + N/A
* How did you select features?
  + I selected features used for visualisation that I would think would tell the story that I was trying to tell. As far as the feature used for modelling, it was the only way I could use the data to answer the business question
* What feature engineering techniques are used?
  + Replacing null values
* What are the models used?
  + Keras LSTM
* How long does it take to train your model?
  + 5-10mins (at 100 epochs)
* What are the tools used? (cloud platform, for example)
  + Nil
* What are the model performance metrics?
  + RMSE
* Which model was selected?
  + LSTM

## Outcomes

* What are the main findings and conclusions of the data science process?
  + The weekend produces higher amounts of fatalities; December is the highest month for fatalities and the results of the LSTM modelling show that through 2021-2022 (May – May,) there is expected to be approximately 90 road fatalities per month. This guided the target value and the recommendation.

## Implementation

* What are the considerations for implementing the model in production?
  + I have only run the model on a 12 month window, with the future predictions outputting 12 months of data, due to the RMSE evaluation metric applied on the validation data. It can be changed for a window (LSTM time-steps) of any amount of months to output for a longer ranging look at predicting future values.

# Data answer

* Was the data question answered satisfactorily?
  + I don’t believe so – I think the data is potentially too simple to be able to accurately reflect the relationship of timeseries – fatalities. I would have liked to include some of the data from the 52,397 features but I wasn’t able to transform the data with enough time to add it into the model.
* What is the confidence level in the data answer?
  + Using the RMSE value as the evaluation metric, the model is approximately 85% accurate. I feel there is a limitation of the output due to the simplicity of the data, as mentioned above.

# Business answer

* Was the business question answered satisfactorily?
  + Yes, the business question was answered satisfactorily – the potential target for the reduction of the road toll would provide a substantial cost saving for governments and transport/freighting businesses.
* What is the confidence level in the business answer?
  + Again, there is no mathematical confirmation of the confidence level of the business answer but I am fairly confident that

# Response to stakeholders

* What are the overall messages and recommendations to the stakeholders?
  + A driver education campaign would be the suggested modality: focusing on being courteous and cautious around peak traffic times and weekends (potentially targeted towards males due to the large number of male fatalities)

# End-to-end solution

* What is the overall end-to-end solution to use the model developed in the project?



# References

* Where are the data and code used in the project? (show a simplified list of main items: notebooks, datasets, exported models)
  + **NOTEBOOKS**
    - MVA\_Capstone (first notebook wrangling raw data – the size of this notebook became too big for my computer to load so I decided to split it across several notebooks to keep file size smaller. Also features most of the visualisation)
    - mva\_capstone\_pp (Notebook with pre-processing)
    - mva\_model (Notebook with modelling)
    - mva\_lstm (
    - mva\_capstone\_univariate\_lstm\_final
  + **DATA**
    - ardd\_fatalities\_apr\_2021.csv (raw data)
    - df\_pp.csv (pre-processed data)
    - df\_my.csv (dataframe of month-year values)
    - df\_model.csv (modelling data to validate LSTM models)
    - df\_lstm.csv (modelling data to predict following 12 months of values [using machinelearningmastery as a guide]
* What are the resources used in the project? (libraries, algorithms, etc)
  + Numpy, Pandas, matplotlib.pyplot, Seaborn, ScikitLearn, Keras LSTM