**Capstone Project**

**Documentation**

**SEVERE ACCIDENTS PREDCITION**

**By Winnie Wetthasinghe (June 2021)**

Table of Content

# Process overview

# Introduction to the

# 2.1 Problem statement

# Problem statement

# Stakeholders

# Business question

# Data question

# Data

# Data science process

## 8.1 Data analysis

## 8.2 Model Implementation.

## 8.3 Outcomes

1. Data answer
2. Business answer
3. Response to stakeholders
4. End-to-end solution
5. References

# Process overview

Capstone project followed this overall end-to-end process for defining, designing and delivering.



# Introduction to the problem statement

According to the World Health Organization, traffic accidents are becoming one of the dominating causes of deaths and injuries around the world. In 2015, there were more than 12.5 million deaths caused by road accidents. Between 20 – 50 million people were injured. This is a global proportion of 18 deaths per $100,000 inhabitant. This has not only become a major cause of road mortality but also a public health problem as well as socio-economic development issue.

It is necessary to address this issue to reduce fatality and implement road safety strategies to establish secure mobility environments.

# 2.1 Problem statement

There are two business problems that can be identified in this domain context. Firstly, to find ways to reduce/control road mortality and injury and secondly, what are the contributing factors that led to those severe accidents?

The project is focusing on finding the contributing factors to the severity level of road accidents and possibly helping to mitigate/ eliminate such occurrences and to develop road safety strategies. This also provides immensely helpful insights for car insurance companies to predict risk levels of a certain cover and what factors to focus on when estimating a quote. Investigating these factors further to be aware of their correlations to an accident being severe, will help when handling claims that are related to road accidents. Insurance companies being aware of these factors will assist them in implementing strategies to ensure that accident-related claims.

This problem is valuable to address because of the issues that emerge from not making effort to reduce the situation. It is not only about reducing road mortality but help reduce public health problems and decrease issues in socio-economic development. This way we can induce benefits to the community and Police departments to implement correction/safe strategies and on a business perspective to assist insurance companies to evaluate their risks and estimate quotes.

Generally, road safety projects including Police departments are exercising certain measures to determine the severity of accidents to minimise them from happening. They gather data at an event of an accident and measure severity based on different indicators.

High Threat to Life Indicator (HTTL): An alternative dimension of severity based on a person’s probability of survival. This metric is based on a person’s worst injury where the lowest SRR (Survival Risk Ratio) of all diagnosis codes for the first admission is used to calculate ICISS (ICD-based Injury Severity Score). ICISS values are banded into two categories and have the following survival probabilities: Yes - at most 94.1% No – at least 94.1%

Injury severity category: A dimension of severity based on a person’s probability of survival. This metric is based on a person’s worst injury where the lowest SRR (Survival Risk Ratio) of all diagnosis codes for the first admission is used to calculate ICISS (ICD-based Injury Severity Score). ICISS values are banded into four categories and have the following survival probabilities: Maximum severity - at most 85.4% High severity - between 85.4% and 96.5% Moderate severity - between 96.5% and 99.2% Minimum severity - at least 99.2%

After the assessment, they focus on the contributing factors that make a certain accident severe in nature through quality analysis and in terms of cause and effect. This cause-and-effect measurement can only indicate so much. At a field of fast pacing and constantly changing life situations, transportation, infrastructure growth, emergency services, climate change and other facets combining make impacts on the causes for a road accident to occur and to keep up with every detail to determine severe accidents prevention better yet preventing them altogether is a challenge. This is where Data Science Strategies come in handy and with the help of Machine Learning models, we can keep the pace and maintain the accuracy of finding such measures that are so critical and time sensitive.

Current measures that are been practiced by Insurance companies are indeed effective and helpful in analysing risks and quote estimations. When the data gets too large and situations change so quickly, this task become difficult to handle or to predict correct measures whilst relying on conventional methods. So, implementing ML, Deep Learning models will help in the long run to ensure the predictions of models will determine the most contributing factors to cause accidents and their mortality.

The desired state in this situation is where the blueprint of this method can be implemented anywhere in a system where the likelihood of a severe accident being happening can be controlled by foreseeing its contributing factors.

It is fair to note that there are several overseas researches out there that have applied similar Data Science methods on this topic. The ways they’ve conducted their study were different and choices of different target variables and methodologies are distinctive to each other. But ML algorithms are common. There is a significant amount of usage of deep learning techniques have been used in those studies. Those researches may have paved the way to address accident causing issues that unique to that particular domain context but not necessarily to what’s been discussed in this project. [1] [2]

# Industry/ domain

The main target fields, industries/domains of this project are Infrastructure & Transport departments, Traffic control Police departments, Car Insurance, Emergency Departments and other research health departments.

To highlight the current state of the above-mentioned industries, they are using conventional methods to handle the causes of severe accidents but there are some recent researches that mainly focusing on solving this issue using Data Science capabilities similar to what’s been done in this project.

The overall industry value chain is firstly Collect data –-> Investigate/study data –-> Implement Models -🡪 Determine features/factors -🡪 incorporate strategies according to the added business values.

Furthermore, the key main concepts/area of focus in these industries are to establish road safety strategies with safe mobile environment, eliminating risks for insurers.

This project benefits not only the afore mentioned industries but also sub categorical industries that are relevant such as Research companies, Emergency hospital services (To provide prompt emergency services at a critical event), Automobile companies (introducing new generation accident-free self-driving cars or implementing more safety features in a vehicle)

# Stakeholders

As mentioned earlier, the stakeholders who would benefit from the outcome of this project are State Police Departments, Car Insurance companies, Research companies, Emergency hospital services.

The main obvious reason that this is something the stakeholders should consider integrating this solution into their current systems is that it will help to mitigate accidents and to establish more safety strategies and practices. This also applies to car insurance stakeholders to reduce insurance risks. Gaining insights coming from scientifically proven methods will further concrete their knowledge that will be immensely beneficial to implement highly effective systems.

It should be noted that the stakeholders’ expectations on this is to have a fast-operating model that gives out accurate measures of predicting severity of potential accidents causes to avoid certain accidents being happening in the future.

# Business question

The main business question can be stated as follows. How can we prevent severe accidents being happening? Is there a way to minimise them?

How can we reduce unnecessary insurance claims and risks? How prepared are we for the unexpected?

Even though it’s not been measured, it is believed that after answering this business question, we can expect a certain reduction in severe accidents percentage that bring mortality, notable reduction in mortality rate, better working safety strategies that align well with public and business benefits and certain decrement in insurance claims percentage that directly meet critical business needs.

# Data question

A correctly defined data question will always help to ensure that the solution has correctly been addressed to the business problem. In this project, we determine what is the optimum implemented model that best describes the severity of accidents to determine the most contributing factors.

The data required to answer this question is that it should contain already measured severity values from historical data. The historical data also should contain information about factors that contributed to the accidents with level of degree of their involvement such as number of young drivers etc. Higher the value, higher the effect it has on the occurred event.

# 7.Data

The dataset selected for this project is about Victorian State Crash Statistics from year 2013 to 2019. The origin of the dataset is from Vicroads.gov.au [3] and the data used for this project is acquired from Kaggle.com [4]. The volume of this dataset is 74908 rows × 63 columns. The dataset is considerably reliable as those are compared with the original dataset. The quality of the raw dataset is considerably good as there weren’t many missing/invalid or null values. The origin of the dataset is a government dataset that are generated, collected, preserved, stored and made publicly available by government entities or for those who are delegated to exercise functions of control, execution or reporting or information concerning road accidents.

Things to consider with the acquired dataset is that Government data is generally available for public for technical support or other aspects but that can also mean that not all the variables of the data set maybe available for public access. These can be gender details, ethnicity and specific demographic data that may not be incompliance with the information privacy laws current in the state.

# A picture containing table Description automatically generated

Figure 1: Screenshot of the dataset

# Text Description automatically generated

Figure 2: All the columns in the dataset

# 8.Data science process

## 8.1 Data analysis

After reading the data, we carried out an investigation of the data including data pre-processing where we checked for data types (converted types if needed-combining ACCIDENT\_DATE and ACCIDENT\_TIME to DATE\_TIME and converting it to Datetime type), check for ranges through finding its describe() method and checked for null values.

After doing these steps replaced null values through data imputation and added new columns with imputed data. After doing all the missing values handling, transformation of variables and cleaning the dataset to make it ready for EDA, we’ve exported it for later usage as Cleaned Victoria Crash Statistics.csv.

For the Exploratory Data Analysis, we’ve considered the timeseries data to plot number of accidents occurred annually over the span of 7 years from year 2013 to 2019. With that we’ve plotted the graph of annual number of accidents occurred over the years due to collision with vehicles. Then plotted number of accidents occurred monthly over the span of 7 years. Then in association with that plotted Number of accidents occurred monthly over the years due to collision with vehicles. Afterwards, we acquired the plot related to the cumber of accident occurred with a day (within 24 hours/hourly based) over the 7 years. We were able gain interesting insights from the plots.

Some insights from plots: - (Same explanations can be referred in the Jupyter NoteBook for Capstone EDA)

Accident’s frequency of over the years-

Chart, bar chart

Description automatically generated

Figure 3: Accidents frequency over the years

It seems year 2019 does not had many accidents reported. Year 2013 also significantly shows a smaller number of accidents compared to other years (2014, 2015, 2016, 2017 & 2018). By the looks of it the number of accidents occurred during 2014-2016 is close but from year 2017 the occurrences reduced. Could it be a reason within those year, road safety methods were applied to experience a significant reduce in accidents? The offset between 2013-2014 & the sudden drop in 2018-2019 caused because the data was not completed for those two years. Roughly they have data only for 3 months of the whole year.

Accident’s frequency of every month over the years-

Chart, bar chart

Description automatically generated

Figure 4: Accidents frequency of every month over the years

From year 2013-2019 the months that had most accidents are October, November, December, Feb, March, July, and August. Highest number of accidents occurred during seasons of holidays such as summer.

Hourly based accidents frequency-

Chart, bar chart, histogram

Description automatically generated

Figure 5: Hourly based accident frequency

During the day highest accidents could be recorded during 3 PM- 5 PM in the evening and followed by morning at 8 AM. It is evident that accidents are high when there is more traffic. So, during the rush hours more accidents were recorded.

Count of accidents per Day of the Week-

Chart, bar chart

Description automatically generated

Figure 6: Number of accidents per day of the week over the years

From the plot it is clear that during Friday, that is the beginning of the weekend there were many accidents reported as it well could be that Friday usually is a busy day for any city country wide specially during the evening hours. Sunday reported to have less accidents there weren't many people on the road.

Accident types occurred over the years-

A picture containing chart

Description automatically generated

Figure 7: Accident types occurred over the years

From the plot above, it seems that collision with vehicles is the main type of accident reported in whole Victoria state. The collision with a fixed object is the second highest reported.

Light condition when accidents occurred over the years-

Chart

Description automatically generated

Figure 8: Light conditions when accidents occurred over the years

Interestingly, most of the accidents occurred during the daytime. Which implies that it could well happened when there was a lot of traffic on the road during the daytime. When people are active on the road.

Road geometry of the places when accidents occurred-

Table

Description automatically generated with medium confidence

Figure 9; Road geometry of places when accidents occurred

Most of the accidents have occurred non intersection areas. Could this mean that at intersection, there are more traffic control happening whereas in non-intersect areas are prone to accidents.

Speed limit of the area when accidents occurred-

Table

Description automatically generated

Figure 10: Speed limit of the areas when accident occurred

It seems most of the cases accidents have happened at 60km/hr speed limit. And Second highest is 50km/hr. Could this mean these accidents occur due to speeding in low-speed zones, especially in Melbourne urban areas?

There are more graphs that are plotted to gain insights and for further information, refer the EDA part of the notebook.

Note:-

In-order to ensure the reusability of the codes/pipeline, we have separated the exercise into two notebooks. One containing the completed EDA part and the other contain the

Completed modelling part where results are derived.

8.2 Model Implementation.

For modelling the target-variable we used is “severity” and for predictive features we’ve selected many features and discarded a few. This is the list of discarded features and used the remaining as –

'SEVERITY', 'ABS\_CODE', 'DCA\_CODE','NODE\_ID', 'OBJECTID', 'LGA\_NAME\_ALL','POLICE\_ATTEND', 'INJ\_OR\_FATAL', 'FATALITY','SERIOUSINJURY','OTHERINJURY', 'NONINJURED','DEG\_URBAN\_ALL','DIVIDED\_ALL\_Imputed','RMA\_ALL\_Imputed','REGION\_NAME\_ALL\_Imputed','ACCIDENT\_DATE', 'ACCIDENT\_TIME

There wasn’t a notable correlation between features. That is one of the reasons to include many predictive features for modelling by producing dummy variables for the process of feature engineering. The second reason is to convert all the categorical data into numerical values.

There are some features that came out as important after performing the feature importance of the selected model XgBoost that got a significant portion of the final result. Those are: -

* Accident Type- Collision with vehicle
* Motorcycle
* Urban Melbourne Areas-Rural Victoria
* Alcohol related
* Old pedestrians
* Old drivers
* Light conditions

After performing feature selection and feature engineering, we implemented several numbers of Machine Learning models to predict the severity level of an accident.

The models are: -

Logistic Regression, Logistic Regression with hyperparameter tunning, Naive Bayes, Decision Tree Classifier, Random Forest Classifier, Random Forest with Class Weighting, KNN Classifier, Bagging Meta Estimator pairing with KNN, Default Bagging Classifier, Gradient Boosting Classifier, Adaboost Classifier, XGBoost Classifier, and Stacking Classifier.

The initial model training time for all the models took a matter of few minutes but training after hyperparameter tuning took considerable amount of time for some models but still less than 10 mins etc. The hyperparameter tuning done with GridSearchCV took more than 15 mins to run the model. The more hyper parameters added, the more time it took to train the models.

The tools used for modelling are all Machine Learning related Python library packages.

The initial performance metrices selected to evaluate the ML models were the Accuracy score and the Recall. The Accuracy scores were relatively acceptable, but the Recall scores were not acceptable. Recall is a crucial performance evaluation metric for classification models that predicts the True Positives or that is getting the probability of a model predicting true values. If this scores poorly, then the considering just the Accuracy level would be void. So, the safe performance metric was AUC Score that give us the rate between False Positive Rate vs True Positive Rate. The higher the AUC score (closer to 1) the better it predicts the 0s being 0s and 1s being 1s. From the ROC curve the models could be compared and any model that has a ROC curve below the random model and give less than 0.5 AUC score is better performing model that can predict the expected target being true. But anything above the 45-degree diagonal line is acceptable and the closer it’s to the top left corner of the plot, the better the model performance is.

The required accuracy or the performance is above 0.8 - 0.85 but the best performing model gave about 0.71 performance even after some hyper-parameter tuning. This further needs to be improved through more hyper-parameter tuning tasks.

The implications of having the model going wrong or getting more values of false positives and false negatives are follows.

The false positive in this case is that the model predicting an accident being severe when actually the accident is non-severe. This means action taken with this result is a waste of effort & use of resource and missing out on addressing the issue. But the implications will only be critical to some extent, and it is will not be fatal. However, the false negative case is different, and it can bring detrimental impacts. Because FN is when the model is not predicting an accident being severe or predicting an accident being non-sever when it is actually severe. Missing this crucial information could miss the chance of preventing a situation being fatal. The worst-case scenario is loss of lives. It is imperative to be attentive to this measure and minimise the occurrence of these unfortunate events.

From the AUC score for performance, we selected three models out of the ones used. But one in particular selected to derive feature importance to determine the most contributing factors to cause severe accidents. The ML model of interest is Xgboost. The reason for this selection will be further explained from the results table.

Table 1: Model evaluation performance metrices

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ML Model** | **Accuracy** | **Recall** | **Precision** | **AUC** |
| Logistic Regression with hyper parameter tuning | 0.65 | 0.63 | 0.44 | 0.70 |
| Naive Bayes | 0.68 | 0.076 | 0.39 | 0.62 |
| Decision Tree Classifier | 0.70 | 0.16 | 0.56 | 0.66 |
| Random Forest Classifier | 0.70 | 0.24 | 0.52 | 0.68 |
| Random Forest with Class Weighting | 0.63 | 0.64 | 0.42 | 0.69 |
| KNN Classifier | 0.64 | 0.24 | 0.53 | 0.55 |
| Bagging Meta Estimator pairing with KNN | 0.64 | 0.27 | 0.37 | 0.51 |
| Default Bagging Classifier | 0.70 | - | - | 0.56 |
| Gradient Boosting Classifier | 0.71 | 0.15 | 0.61 | 0.70 |
| AdaBoost Classifier | 0.71 | 0.20 | 0.57 | 0.69 |
| XGBoost Classifier | 0.71 | 0.26 | 0.55 | 0.70 |
| Stacking | 0.60 | 0.32 | - | - |

Considering the results above, we’ve selected XGBoost as the model to find feature importance as the current result for this project. But these results can be improved further and that purpose we’ve selected 3 most high performing models such as XGBoost, AdaBoost and Linear Regression with particular hyper-parameters tuned.

## 8.3 Outcomes

The key findings and conclusions from this project are as follows. Please note that these findings can be changed upon doing more future tasks on this project to improve scores of model performance.

1. Logistic Regression, Gradient Boost and XGBoost Classifiers performed well (These can be used further). Even at the initial modelling stage XGBoost performed well and it is still a powerful ML classifier.
2. Proper hyper-parameter tuning can improve model performance.
3. The outcomes can be immensely improving with more hyper-parameter tuning and diligent feature engineering. The features in the dataset are not entirely sufficient to predict severity. A strong severity predictor can be implemented using more highly correlating features.
4. The XGBoost model was overfitting with the training data when the max-depth exceeds its optimum value which is 3 (Max\_depth = 3)

Chart, line chart

Description automatically generated

Figure 11: Over fitting of the train data and test data

1. The GridSearchCV hyper-parameter tuning proved that max\_depth=3 is the optimum number and gave optimum values for gamma as 0.5 and accuracy as 0.714.

A picture containing text

Description automatically generated

1. Two most crucial contributing factors that are affecting the severity level of accidents are collision with another vehicle and accidents involved in motorcycles.
2. The next important features that need to be considered are accidents happening in Melbourne urban areas, alcohol related accidents, safety of elderly pedestrians, elderly driver, and light conditions.

# 

# 9.Data answer

The data question that was addressed is in acceptance but not entirely in satisfactory. The reason because is that the model can be tuned to perform better. But for now, selection of XGBoost is the most optimum. After doing the reverse engineering work to find the important factors (features), the results are in satisfactory. Some of them were anticipated and others are interesting findings.

We acquired the confidence interval by assuming of a Gaussian distribution of the proportion and calculated the radius of the confidence interval.

The classification error of the model is 36% +/- 7% The true classification error of the model is likely between 29% and 43%.

# 10.Business answer

The business question was How can we reduce road mortality and injury? And what factors have most effect on mortality? The way we’ve designed the solution, the best approach to reduce road mortality is done by addressing high contributing factors for the causes of severe accidents. So, once we find those factors then the business stakeholders can utilise the results and pay more attention to the factors by addressing them in ways that’s necessary and feasible. From the project we derived a list of critical factors that need to be considered when implementing solutions. The business answer is satisfactory with the current results, but this too can be improved.

# 11.Response to stakeholders

The overall message and the recommendations to the stakeholders who will be benefiting from this project results are as follows.

* Utilise and apply the problem-solution approach into prevailing problems. State Police Department can use these insights to implement and improve road safety strategies that help prevent severe accidents.
* Take these insights into consideration when applying business strategies. Car insurance companies can use this information to predict expected insurance claims and their nature of things and then incorporate those strategies to minimise risks.
* The solution model is currently performing at an acceptable level but needs more improving in order to get more accurate output with better performing models. Perform more hyperparameter tunning for the selected models and feature engineering work to increase scores.
* Apply deep learning techniques to further implementations of models
* Incorporate NLP to detect recent traffic information such as social media related & open data sources to gather data.

# 

# 12.End-to-end solution

This is the overall end-to-end solution to use the model developed in the project. This basically discuss about applying the implemented model into data pipeline using data engineering techniques.

After deploying our machine learning model to a pipeline these steps need to be considered.

INVESTIAGTE THE BUSINESS & DATA QUESTION RELEVANCY

Is the business question changing? Keep investigating whether this model needs re-modelling?

Is the mortality rate is changing? Keep testing with new data and see whether the current model can keep up with the changing conditions or is it changing at all?

DATA MINING & INGESTING DATA

Take batch wise or streaming data for re-modelling. (Receiving data from end user devices 🡪 polling proxy and queuing🡪 create subscriptions via pub/sub methods to send it to the data processor)

DATA STORAGE & TRANSFORMATION

Regularly store data in secure data storages (Cloud storages, BigQuery etc) Transform data and. Import processed data for reuse (modelling purposes).

DATA PROCESS AND ANALYSE

This is where our ML models run for modelling. Data scientists mainly present here to monitor model performance. Further data exploration will be done here as well.

# 13.References

**Codes**

Exploratory Data Analysis Notebook –

<https://github.com/Winnie-Sepz/Data-Science-Course/blob/main/Data%20Science%20Course%20UTS/Capstone%20project-WInnie%20Wetthasinghe/Capstone%20Project%20Part%201-%20EDA%20-Winnie%20Wetthasinghe.ipynb>

Implementation of Machine Learning Models Notebook-

<https://github.com/Winnie-Sepz/Data-Science-Course/blob/main/Data%20Science%20Course%20UTS/Capstone%20project-WInnie%20Wetthasinghe/Capstone%20Project%20Part%202%20-%20Modelling%20-Winnie%20Wetthasinghe.ipynb>

Victoria Crash Statistices.csv - <https://drive.google.com/drive/u/0/folders/1qm2uZzug_lNe0A5sTmaoleHXaqzp7Ubn>

Cleaned Victoria Crash Statistics.csv. - <https://drive.google.com/drive/u/0/folders/1qm2uZzug_lNe0A5sTmaoleHXaqzp7Ubn>

**Resources**

[1]<https://reader.elsevier.com/reader/sd/pii/S209575642030101X?token=3E750E36FF75A24665601D09A2597DC8572E83CDEC34DC69480A5D842B29DADE5E21437CB57FA76F0BDB39F57866F828&originRegion=us-east-1&originCreation=20210530224527>

[2]<https://www.researchgate.net/publication/220166391_Traffic_Accident_Analysis_Using_Machine_Learning_Paradigms>

[3]<https://data.vicroads.vic.gov.au/metadata/Attribute_Table_Viewlist_7ac33a09.asp.html>

[4] <https://www.kaggle.com/gaurav896/victoria-state-accident-dataset>

[5] <https://github.com/NFaraji/An-analysis-of-Victoria-crash-data/blob/master/Insight-Crashes.ipynb>

[6] <https://data.vicroads.vic.gov.au/metadata/crashstats_user_guide_and_appendices.pdf>