IBM Data Science Specialisation

Capstone Project

Predicting Automobile Crash Severity in New Zealand

Fernando Castellanos

October 2020

Business Problem/ Problem Description and Background

- World Health Organization, road traffic injuries caused an estimated 1.35 million deaths worldwide in the year 2016

  (Global status report on road safety 2018, WHO)

- More than half of all the deaths are among vulnerable users: cyclists, pedestrians or motorcyclists.
- Road traffic injuries are currently the leading cause of death for children and young adults.
- Even with all the safety devices installed in modern cars (such as seat belts, airbags, anti-lock brakes, shatter-resistant glass and head restraints), fatalities due to automobile crashes still occurring in high numbers

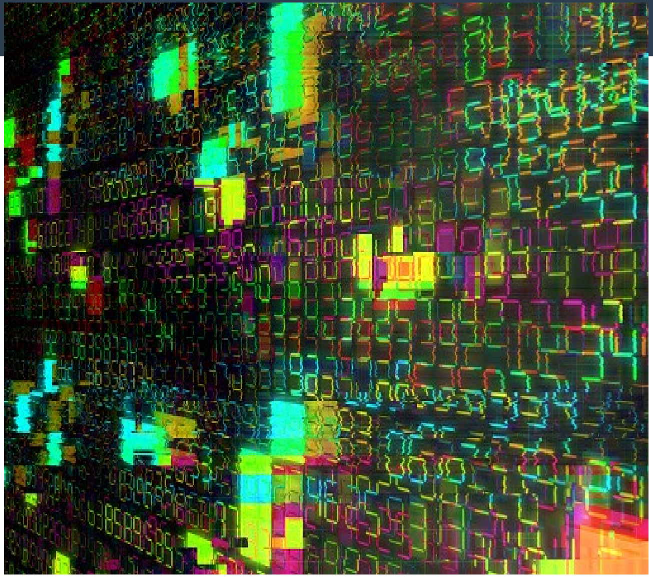
Target Audience & Deliverables

- Data collection and analysis allow to better understand how automobile crashes happen and how to attempt to predict their outcome (severity).
- Possibility to enable the transport, security and emergency local agencies to analyse incoming reports, determine on-line the crash severity and dispatch the adequate emergency response and activate any traffic controls.
- Nowadays, transport agencies around the world collect data of all reported automobile crashes.
- The goal is to predict the severity of automobile crash samples.
- Project files are available on my Github repository here, except data (file too big). Main files are: jupyter notebook (Capstone_project_crashes_final.ipynb0, project report (Capstone_final_report) and this presentation.

The Dataset

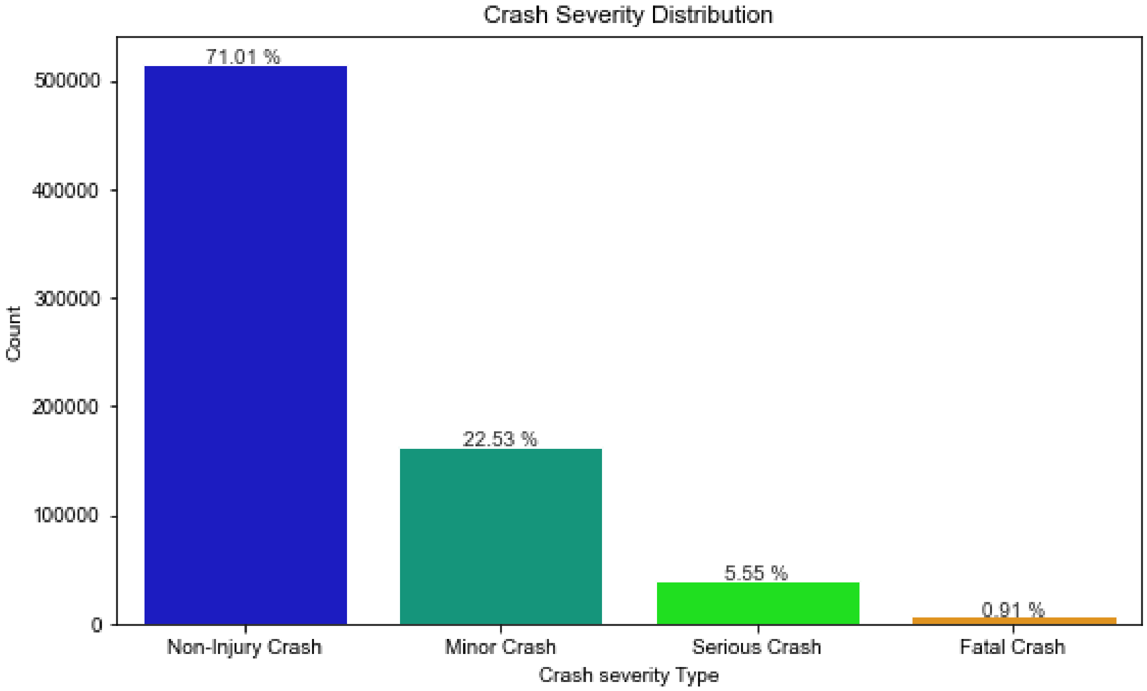
- New Zealand Transport Agency open data repository, *”Crash Analysis System (CAS) data”*
- Data file:
  - Crashes reported since 1 January 2000
  - Updated quarterly, last time July 2020
  - 725548 samples and 72 features
  - Description features on Appendix A, project report
  - Target feature is *‘crashSeverity’*. It has four possible values: *‘Fatal Crash’*, *‘Serious Crash’*, *‘Minor Crash’* and *‘Non-Injury Crash’*

The Dataset



- Features in the dataset are categorical, numerical and geodata.
- Categorical features provide information about different attributes and conditions of the crash event, e.g. “*roadLane*” informs about the road configuration.
- Numerical features are mainly counters providing details about a variety of objects that could be involved in the crash, e.g. “*parkedVehicle*” indicates how many times a parked vehicle was struck at the crash site.
- Another set of numerical features provide values for variables such as *“speedLimit”*.

The Target Feature '*crashSeverity*'



- It is a categorical feature, with 725548 samples.
- It has four classes: '*Non-Injury Crash*', '*Minor Crash*', '*Serious Crash*' and '*Fatal Crash*'.
- Large imbalance between the four classes.
- Imbalance is in the nature of the feature.

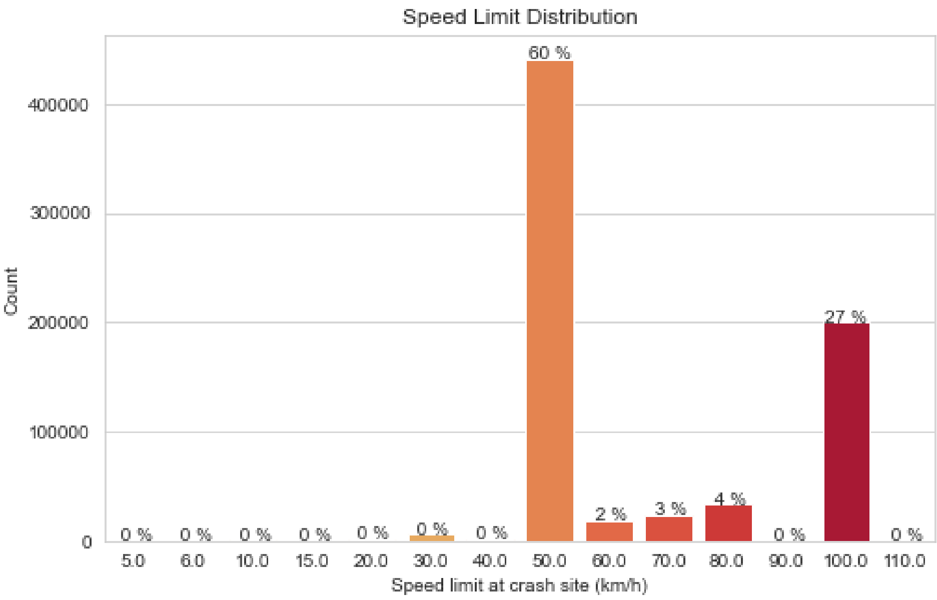
Data Cleaning

- It is a very large dataset with 725548 samples and 72 features.
- Removed 15 features related to geolocation identifiers, unique identifier codes, year of crash, specific location description crash site and number of people injured because they do not provide information relevant to the analysis.
- Removed two more ‘*crashRoadSideRoad*’ and ‘*intersection*’ because, they do not have data.
- Removed two more ‘*advisorySpeed*’ and ‘*temporarySpeedLimit*’ because, they have very few samples.
- Number of features was reduced from 72 to 53.

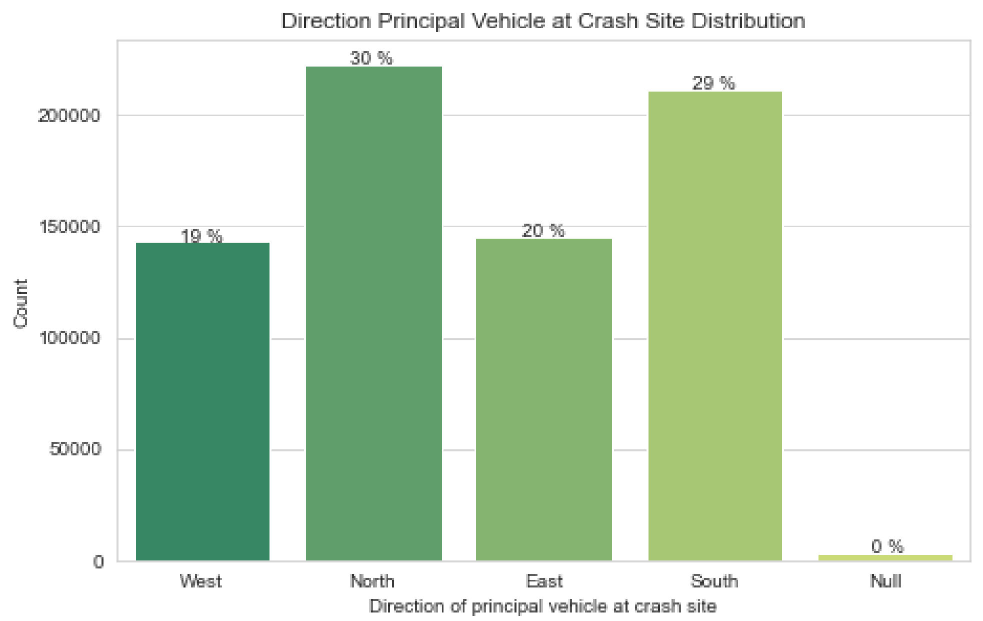
Data Exploration – Missing Data, Sanitation

Some brief examples of the analysis

- Many features have missing values (NaN), the majority are counters. Their NaN values most probably represent the no presence of the features element at the crash. Hence, all these NaN can be replaced by 0.
- *‘speedLimit’* has 492 NaN values, only one of them is *‘Fatal Crash’* and most are *‘Non-Injury Crash’*. NaN values replaced by 50 or 100 km/h depending value another feature ‘urban’.
- ‘tlaName’ has 35 NaN values, most of them are *‘Non-Injury Crash’* and there is none *‘Fatal Crash’*. The NaN samples were deleted.



Data Exploration – Missing Data, Sanitation



Similar analysis and procedures applied to:

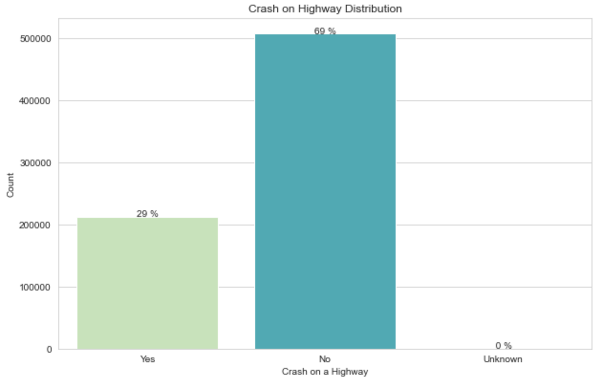
- ‘*NumberOf Lanes*’ (1193 NaN values replaced with ‘2 lanes’)
- ‘*holiday*’ (685925 NaN values replaced with ‘None’)
- ‘*directionRoleDescription*’ (118 samples with ‘NaN’ values and 3070 samples with ‘Null’. Only six are ‘*Fatal Crash*’. All ‘NaN’ and ‘Null’ samples were deleted).

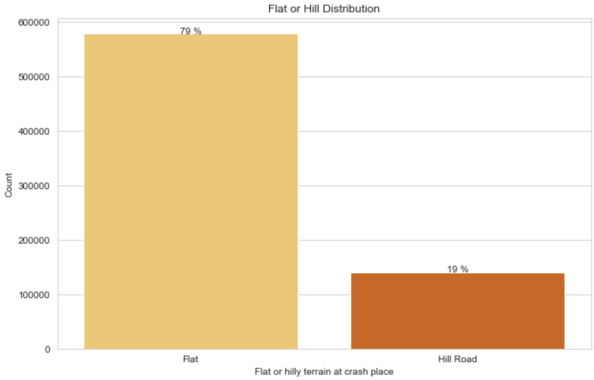
Data Exploration – Review Categorical Features, Sanitation

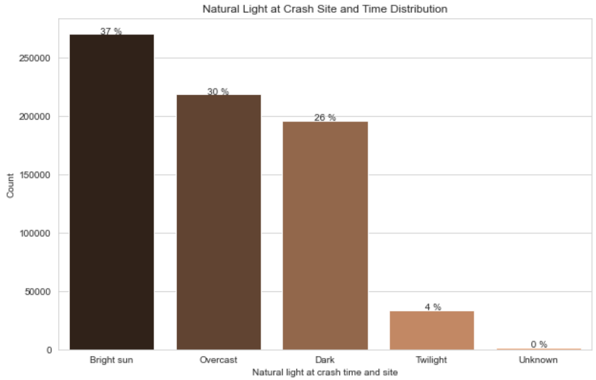
The review provided insights into the categorical features. For each feature the review covered the following main aspects:

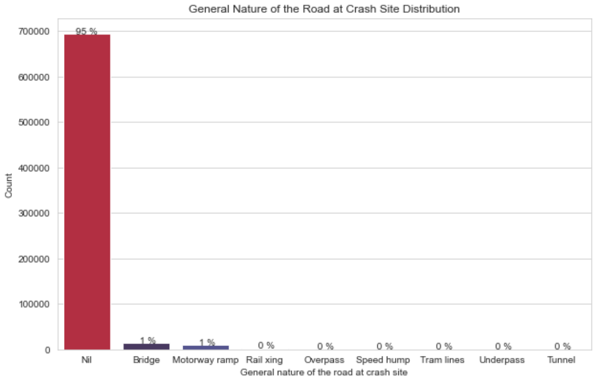
- Analysis of the distribution of values between the different classes.
- To study the feature behaviour (e.g. Main classes and their expected conduct). For example it was found that: 78% crashes occur in fine weather, 68% occur in urban areas, 98% occur on sealed roads, 90% occur on 2-way roads and 30% occur on a highway.
- Compliance of the classes names and the feature definitions provided with the dataset. Any class discrepancy was further studied in terms of its values distribution within the target feature and weight possibilities of renaming it, replacing it or deleting it.
- Plots counts of the classes (the plots did not include NaN values, to make them clearer).

Data Exploration – Review Categorical Features, Sanitation


Crash on Highway Distribution


Flat or Hill Distribution


Natural Light at Crash Site and Time Distribution


General Nature of the Road at Crash Site Distribution

Examples of some features

- *‘crashSHDescription’* (~ 30% crashes occur on a Highway)
- *‘flatHill’* (~ 80% crashes occur on flat terrain. 3115 samples deleted non-existent class)
- *‘light’* (equally distributed three classes: ‘Bright sun’, ‘Overcast’ and ‘Dark’. Samples class ‘Unknown’ are valid)
- *‘roadCharacter’* (96% data is ‘Nil’, is valid. Most crashes at bridges, motorway ramps and rail xings)

Data Wrangling – Conversion and Dataset Summary Changes

Initially numerical data was read as float type and categorical as object. Most numerical were converted to integer type and categorical to category type.

After cleaning, reviewing and wrangling the dataset

	Number of samples	Number of features
Initial size dataset	725548	72
New size dataset	718886	53
Difference	6662	19

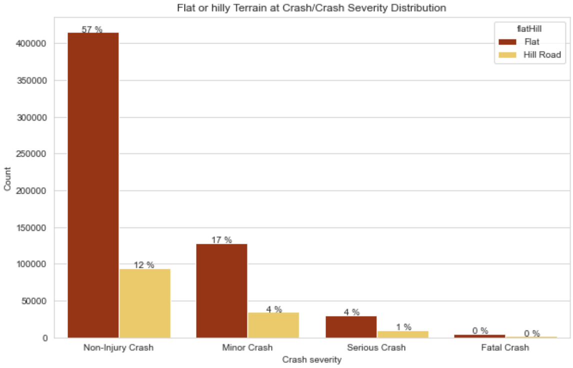
Data Exploration – Features Correlation

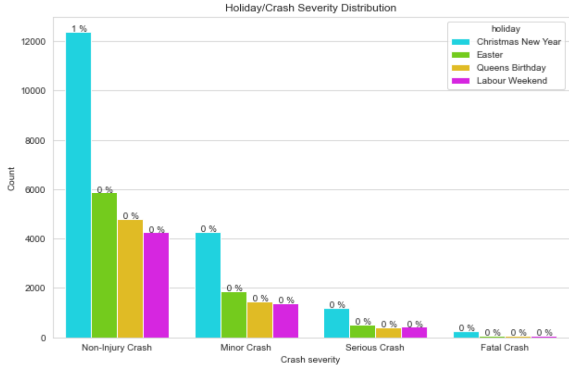
Visual exploration

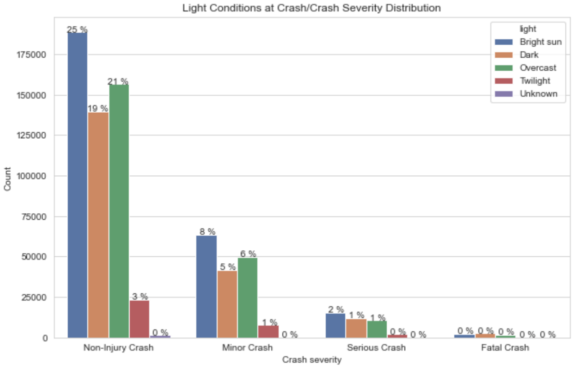
- Plots of all the categorical features
- Plots feature classes against the target feature classes
- There is no clear relation visible in the plots

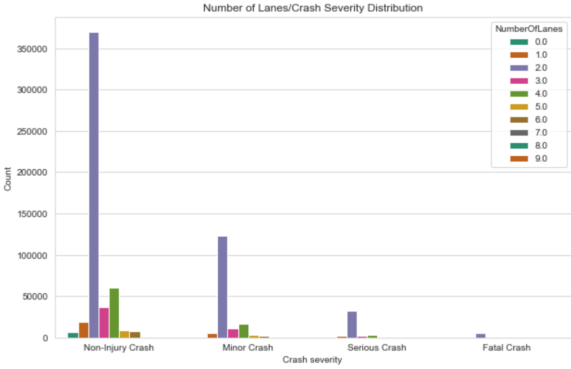
Correlation matrix

- Two new methodologies: phi_K and Dython (Details in project report)
- Chi-square



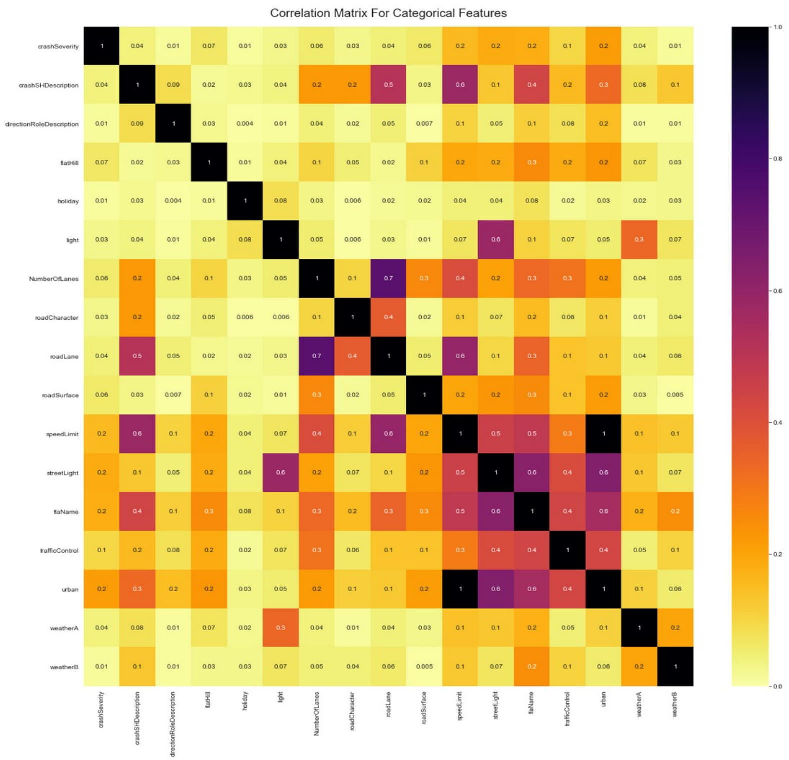






Data Exploration – Correlation Matrix

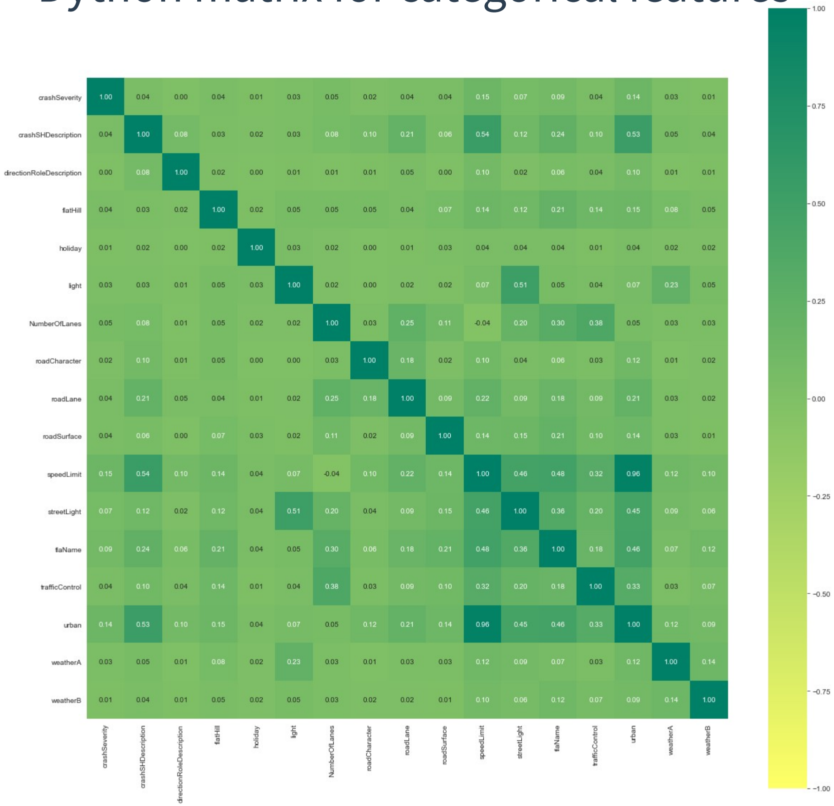
phi_K matrix for categorical features



Correlation Matrix For Categorical Features

	crashSeverity	crashSHDescription	directionRoleDescription	flatHill	holiday	light	NumberOfLanes	roadCharacter	roadLane	roadSurface	speedLimit	streetLight	flaName	trafficControl	urban	weatherA	weatherB
crashSeverity	1	0.04	0.01	0.07	0.01	0.03	0.06	0.03	0.04	0.06	0.2	0.2	0.2	0.1	0.2	0.04	0.01
crashSHDescription	0.04	1	0.09	0.02	0.03	0.04	0.2	0.2	0.5	0.03	0.6	0.1	0.4	0.2	0.3	0.08	0.1
directionRoleDescription	0.01	0.09	1	0.03	0.004	0.01	0.04	0.02	0.05	0.007	0.1	0.05	0.1	0.08	0.2	0.01	0.01
flatHill	0.07	0.02	0.03	1	0.01	0.04	0.1	0.05	0.02	0.1	0.2	0.2	0.3	0.2	0.2	0.07	0.03
holiday	0.01	0.03	0.004	0.01	1	0.08	0.03	0.006	0.02	0.02	0.04	0.04	0.08	0.02	0.03	0.02	0.03
light	0.03	0.04	0.01	0.04	0.08	1	0.05	0.006	0.03	0.01	0.07	0.6	0.1	0.07	0.05	0.3	0.07
NumberOfLanes	0.06	0.2	0.04	0.1	0.03	0.05	1	0.1	0.7	0.3	0.4	0.2	0.3	0.3	0.2	0.04	0.05
roadCharacter	0.03	0.2	0.02	0.05	0.006	0.006	0.1	1	0.4	0.02	0.1	0.07	0.2	0.06	0.1	0.01	0.04
roadLane	0.04	0.5	0.05	0.02	0.02	0.03	0.7	0.4	1	0.05	0.6	0.1	0.3	0.1	0.1	0.04	0.06
roadSurface	0.06	0.03	0.007	0.1	0.02	0.01	0.3	0.02	0.05	1	0.2	0.2	0.3	0.1	0.2	0.03	0.005
speedLimit	0.2	0.6	0.1	0.2	0.04	0.07	0.4	0.1	0.6	0.2	1	0.5	0.5	0.3	1	0.1	0.1
streetLight	0.2	0.1	0.05	0.2	0.04	0.6	0.2	0.07	0.1	0.2	0.5	1	0.6	0.4	0.6	0.1	0.07
flaName	0.2	0.4	0.1	0.3	0.08	0.1	0.3	0.2	0.3	0.3	0.5	0.6	1	0.4	0.6	0.2	0.2
trafficControl	0.1	0.2	0.08	0.2	0.02	0.07	0.3	0.06	0.1	0.1	0.3	0.4	0.4	1	0.4	0.05	0.1
urban	0.2	0.3	0.2	0.2	0.03	0.05	0.2	0.1	0.1	0.2	1	0.6	0.6	0.4	1	0.1	0.06
weatherA	0.04	0.08	0.01	0.07	0.02	0.3	0.04	0.01	0.04	0.03	0.1	0.1	0.2	0.05	0.1	1	0.2
weatherB	0.01	0.1	0.01	0.03	0.03	0.07	0.05	0.04	0.06	0.005	0.1	0.07	0.2	0.1	0.06	0.2	1

Dython matrix for categorical features



	crashSeverity	crashSHDescription	directionRoleDescription	flatHill	holiday	light	NumberOfLanes	roadCharacter	roadLane	roadSurface	speedLimit	streetLight	flaName	trafficControl	urban	weatherA	weatherB
crashSeverity	1.00	0.04	0.00	0.04	0.01	0.03	0.05	0.02	0.04	0.04	0.15	0.07	0.09	0.04	0.14	0.03	0.01
crashSHDescription	0.04	1.00	0.08	0.03	0.02	0.03	0.08	0.10	0.21	0.06	0.54	0.12	0.24	0.10	0.53	0.05	0.04
directionRoleDescription	0.00	0.08	1.00	0.02	0.00	0.01	0.01	0.01	0.05	0.00	0.10	0.02	0.06	0.04	0.10	0.01	0.01
flatHill	0.04	0.03	0.02	1.00	0.02	0.05	0.05	0.05	0.04	0.07	0.14	0.12	0.21	0.14	0.15	0.08	0.05
holiday	0.01	0.02	0.00	0.02	1.00	0.03	0.02	0.00	0.01	0.03	0.04	0.04	0.04	0.01	0.04	0.02	0.02
light	0.03	0.03	0.01	0.05	0.03	1.00	0.02	0.00	0.02	0.02	0.07	0.51	0.05	0.04	0.07	0.23	0.05
NumberOfLanes	0.05	0.08	0.01	0.05	0.02	0.02	1.00	0.03	0.25	0.11	-0.04	0.20	0.30	0.38	0.05	0.03	0.03
roadCharacter	0.02	0.10	0.01	0.05	0.00	0.00	0.03	1.00	0.18	0.02	0.10	0.04	0.06	0.03	0.12	0.01	0.02
roadLane	0.04	0.21	0.05	0.04	0.01	0.02	0.25	0.18	1.00	0.09	0.22	0.09	0.18	0.09	0.21	0.03	0.02
roadSurface	0.04	0.06	0.00	0.07	0.03	0.02	0.11	0.02	0.09	1.00	0.14	0.15	0.21	0.10	0.14	0.03	0.01
speedLimit	0.15	0.54	0.10	0.14	0.04	0.07	-0.04	0.10	0.22	0.14	1.00	0.46	0.48	0.32	0.96	0.12	0.10
streetLight	0.07	0.12	0.02	0.12	0.04	0.51	0.20	0.04	0.09	0.15	0.46	1.00	0.36	0.20	0.45	0.09	0.06
flaName	0.09	0.24	0.06	0.21	0.04	0.05	0.30	0.06	0.18	0.21	0.48	0.36	1.00	0.18	0.46	0.07	0.12
trafficControl	0.04	0.10	0.04	0.14	0.01	0.04	0.38	0.03	0.09	0.10	0.32	0.20	0.18	1.00	0.33	0.03	0.07
urban	0.14	0.53	0.10	0.15	0.04	0.07	0.05	0.12	0.21	0.14	0.96	0.45	0.46	0.33	1.00	0.12	0.09
weatherA	0.03	0.05	0.01	0.08	0.02	0.23	0.03	0.01	0.03	0.03	0.12	0.09	0.07	0.03	0.12	1.00	0.14
weatherB	0.01	0.04	0.01	0.05	0.02	0.05	0.03	0.02	0.02	0.01	0.10	0.06	0.12	0.07	0.09	0.14	1.00

Data Exploration – Correlation Matrix and Feature Selection

The Chi-square Test of Independence was used for the categorical features and the results confirmed the results obtained by the new two methods.

All three methodologies did not show any strong correlation between the target feature and another categorical feature.

Full correlation matrices with all the categorical and numerical (53 features x 53 features) were calculated for the two methodologies (no shown here because the size, refer to the project report). The matrices showed a similar pattern to the categorical ones: there are no strong correlations between the target feature and other features, only some weak ones.

All the methods indicated that there are no good correlations with the target feature. However, using the results a selection of 10 features to proceed with the modelling.

Splitting Train & Test - Imbalanced Dataset

The selected features built the dataset to model with classifiers.
First step after was to split the dataset between training and testing data.

Class	Number of samples (percentage)
Non-Injury Crash	509326 (71.01 %)
Minor Crash	162882 (22.53 %)
Serious Crash	40104 (5.55 %)
Fatal Crash	6574 (0.91 %)

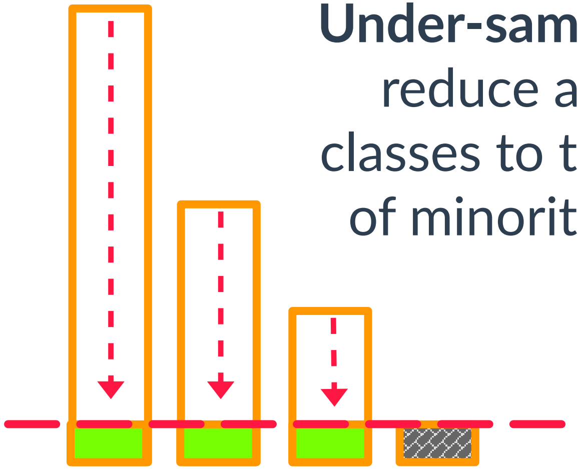
Class distribution for target feature ‘*crashSeverity*’

The target feature ‘*crashSeverity*’ is extremely imbalanced.

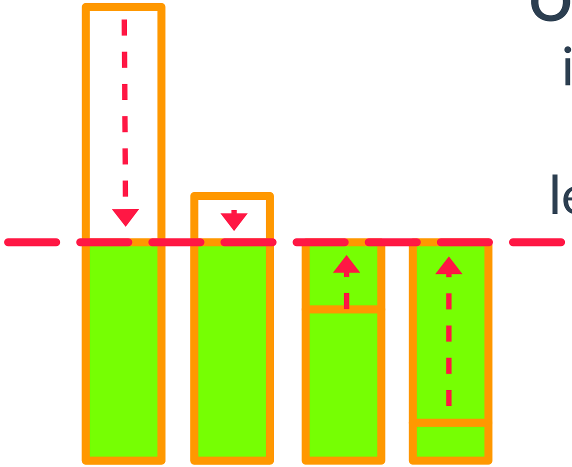
Three ways to manage this:

- Resample the dataset, to artificially balanced it
- Use a variety of classifiers, to avoid bias.
- Use a variety of performance metrics, to avoid bias.

Imbalanced Dataset – Resample



Under-sampled: to reduce all the classes to the level of minority class



Over-sampled: to increase all the classes to the level of majority class

For this project all the classes reduced to 4931 samples.

Because the large difference between the classes in ‘*crashSeverity*’, 50000 samples for each class was selected as the level for the over-sampled dataset.

Three datasets – Encoding Categorical Features

To observe the effects of imbalance and re-sampling, three datasets were used:

- An over-sampled dataset
- The original dataset (named ‘Full set’)
- An under-sampled dataset

Because the majority of classifiers only accept numerical data as input, all the categorical features had to be encoded. The project used ‘OneHotEncoder’. Initially the datasets had 10 columns (10 features), after encoding they finished with 100 columns.

Classifiers/Machine Learning Algorithms

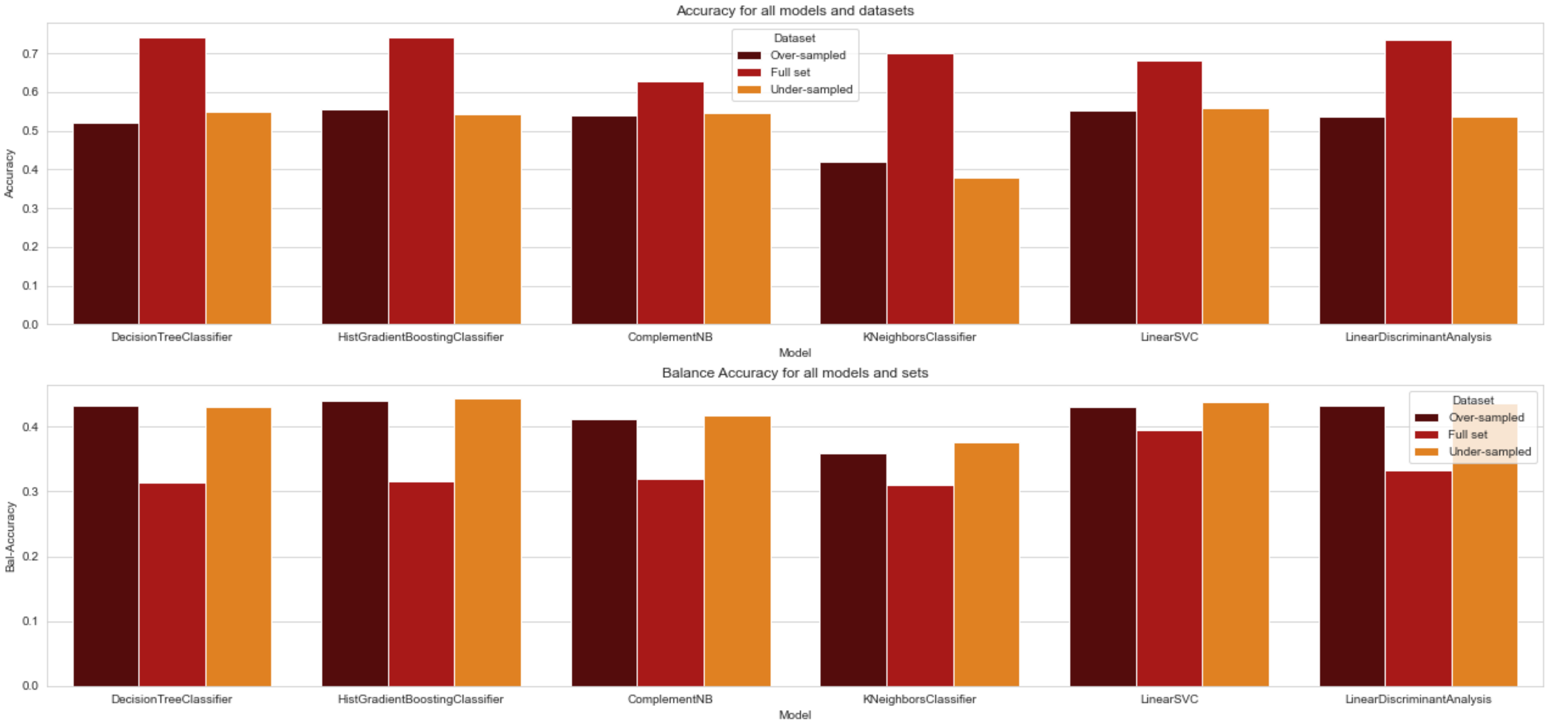
Six classifiers were used for the project:

- Decision Tree Classifier.
- Histogram-based Gradient Boosting Classification Tree
- Complement Naive Bayes Classifier
- k-Nearest Neighbor Classifier
- Linear Support Vector Classifier
- Linear Discriminant Analysis

The performance metrics selected were:

- Precision, recall and f1-score weighted averages
- Accuracy and balanced accuracy
- Precision, recall and f1-score at class level
- Confusion matrix

Results – Average Metrics – Accuracy and Balanced Accuracy



Accuracy best dataset *'Full set'*

Balanced accuracy worst dataset *'Full set'*

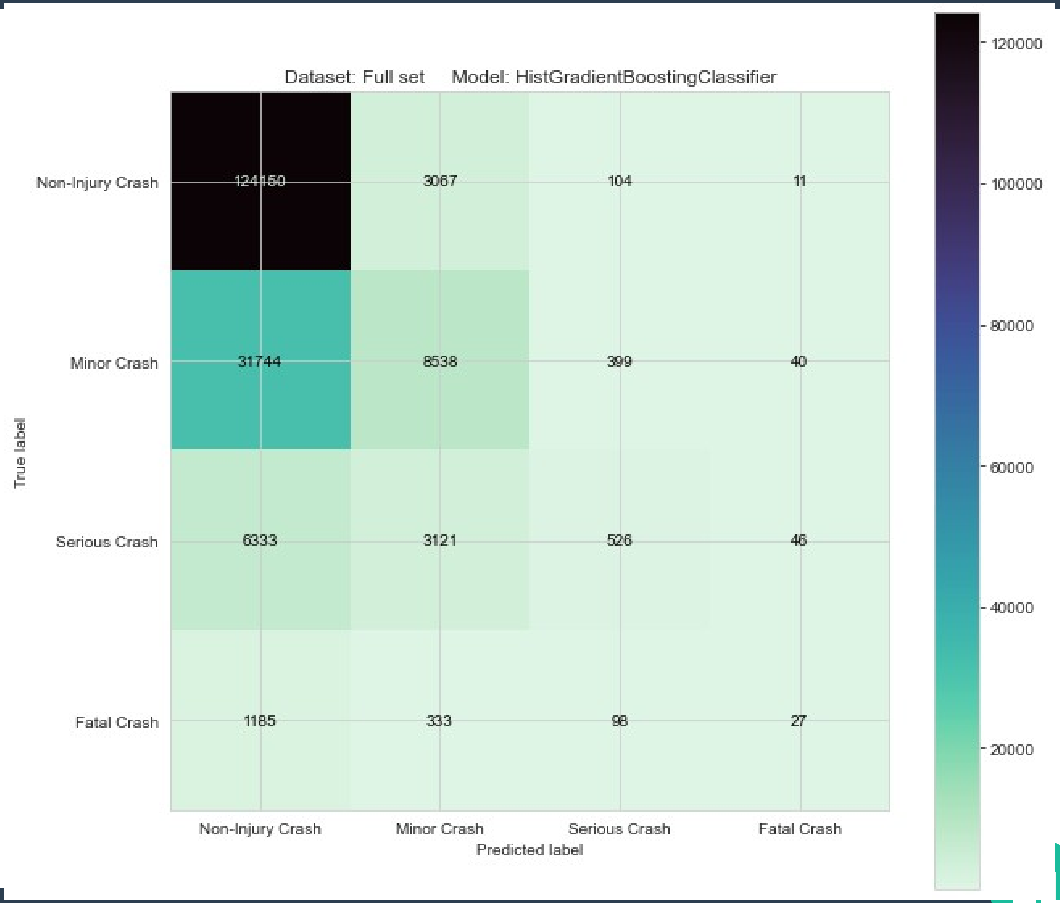
Several classifiers have similar performances. Some classifiers such as k-Nearest Neighbor Classifier appear not to be a good choice for the project.

Results – Average Metrics – Best Combination Dataset and Classifier

Using the average weighted values of precision, recall, f1-score and accuracy it was established that the best combination would be *‘Full set’* with *‘HistGradientBoostingClassifier’*. The confusion matrix for this scenario is shown here.

For *‘Fatal Crash’* and *‘Serious Crash’* only 1.6% and 5.2% are predicted correctly.

A large number of misclassify samples are labelled as *‘Non-Injury Crash’.*

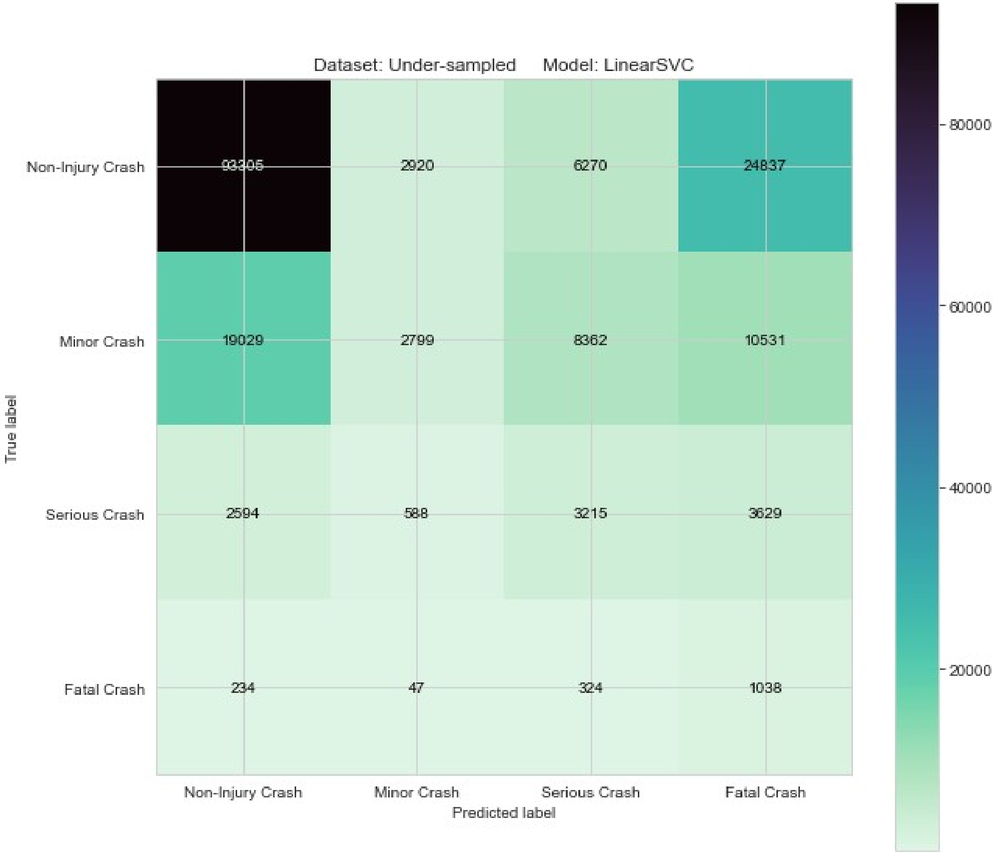


Results – At The Class Level Metrics – Best Combination Dataset and Classifier

Using the class level average values of precision, recall, f1-score it was established that the best combination would be *‘Under-sampled’* with *‘LinearSVC’*.

The confusion matrix for this scenario is shown here.

The correct classification of *‘Fatal Crash’* samples has improved considerably (63%) however, very large number of samples are misclassified as *‘Fatal Crash’.*



Dataset: Under-sampled   Model: LinearSVC

True label \ Predicted label	Non-Injury Crash	Minor Crash	Serious Crash	Fatal Crash
Non-Injury Crash	93305	2920	6270	24837
Minor Crash	19029	2799	8362	10531
Serious Crash	2594	588	3215	3629
Fatal Crash	234	47	324	1038

Results – Comparison of Metrics Between The Two Selected Combinations

	Metrics					
	Precision		Recall		f1-score	
Classes	HistG & Full set	Linear SVC & Under	HistG & Full set	Linear SVC & Under	HistG & Full set	Linear SVC & Under
Non-Injury Crash	0.760	0.810	0.975	0.733	0.854	0.770
Minor Crash	0.569	0.441	0.209	0.069	0.306	0.119
Serious Crash	0.467	0.177	0.058	0.321	0.103	0.228
Fatal Crash	0.252	0.026	0.018	0.632	0.034	0.050

The metrics for *'Non-Injury Crash'* do not change drastically, only some recall and f1-score reductions. For *'Minor Crash'* the values decrease as well, but with larger differences. For *'Serious Crash'* there is a reduction for precision, but increases for recall and f1-score. For *'Fatal Crash'*, there is a sharp decrease on precision, a large increase on recall and a slight increase on f1-score.

These changes support the differences seem in the confusion matrices.

Conclusions

Working in this project has been of great benefit for me, as it helped me to have a better understanding of the different steps of data analysis and how they interact with each other. Some of the main aspects are:

- Work with a mainly categorical dataset, use of encoding with categorical features, methods to analyse correlation between the categorical features and categorical and numerical data.
- Work with a highly unbalanced dataset, use of special methods and techniques to overcome the skewness of the data, such as resampling, metrics at class level.
- Use a variety of plots to observe behaviour, distribution and others.

Conclusions

The initial goal of this project was to train a model that would be able to predict the crash severity, ‘crashSeverity’. Unfortunately, this goal was not achieved even after applying techniques such as resampling, encoding and using several classifiers.

The main reason for the lack of success is that ‘crashSeverity’ does not have strong relations with any of the other features. The selected features were the best ones from a pool of weakly correlated features.

From the analysis two combinations were selected as the best, using metrics:

▻ ‘Full set’ dataset and classifier HistGradientBoostingClassifier (selected using average metrics)

▻ ‘Under-sampled’ and classifier LinearSVC (selected using at class level metrics)

Conclusions

From the results but especially from the confusion matrices is possible to conclude:

‘Full set’ dataset and classifier HistGradientBoostingClassifier

- Accuracy = 0.74155, balanced accuracy = 0.31499
- From confusion matrix. ‘Non-Injury Crash’ recall (% correct prediction) = 97.5%. ‘Fatal Crash’ recall (% correct prediction) = 1.6%
- large amount of misclassifies are labelled as ‘Non-Injury Crash’ from the other classes.
- Very few ‘Serious Crash’ and ‘Fatal Crash’ samples are correctly classified. Most of the misclassifies are labelled as ‘Non-Injury Crash’.
- If in the real world, this classifier would dispatch basic services to most crashes when in reality they require a full emergency services team. This would endanger lives and property

Conclusions

‘Under-sampled’ dataset and classifier LinearSVC

- Balanced accuracy = 0.43849, accuracy = 0.55840
- From the confusion matrix. ‘Non-Injury Crash’ recall (% correct prediction) = 73.3%. ‘Fatal Crash’ (% correct prediction) = 63.2%, almost at the same level that for ‘Non-Injury Crash’
- Many misclassifies are labeled as ‘Fatal Crash’ from the other classes. The misclassifies are many times more than the number of ‘Fatal Crash’ samples
- A good amount of ‘Serious Crash’, ‘Minor Crash’ and ‘Non-Injury Crash’ are correctly classified
- The amount of misclassified samples is several times more than the correctly classified ‘Fatal Crash’.
- If in the real world it would send a full emergency team for minor crashes, this would result in wasting emergency resources, time and money.

Conclusions

- Two new correlation matrices, including categorical and numerical features, were found and proved to deliver good results.
- Resampling proved to help all the classifiers to improve their performance.
- Average metrics do not provide a complete view of the classes and tend to hide the poor performance of minority classes. This is especially important for imbalanced datasets.
- The only average metric that provides some clarity when the dataset is imbalanced is the ‘Balanced Accuracy’.
- At class level metrics provide a complete view of the different classes and allows to recognise any class poor performance.
- Confusion matrix has shown to be the best metric. The matrix allows a very complete review of the way the testing set was classified.