# The Capstone – Predicting Traffic Collision Severity

**This is the final project for the data science course: It is only a testing project, no real results and outputs are generated thus the developed models should not be used for any interpretation!**

## Introduction

### Background


*Fig. 1: https://www.who.int/violence_injury_prevention/road_safety_status/2018/CAR-2.gif?ua=1*

Road traffic crashes are globally among the leading causes of death and even exceed serious diseases such as HIV/AIDS [1]. The number of killed persons due to traffic crashes is estimated to be around 1.35 million while less protected traffic participants such as pedestrians or cyclist are most affected. High-income countries count for high usage of vehicles; however, fatal rate is 3 times lower than in low-income countries but still – morbidity and mortality due to traffic crashes remain high and globally represent a high social, health and economic burden. In the global status report on road safety the WHO states an unacceptably high number of deaths on the world's roads [2]. Governments all over the world have taken several (legislative) measures to minimize behavioural risk factors such as speed and drink-driving, or to improve road infrastructure and vehicles safety, thus decreasing the occurrence of road traffic collisions. At the same time, methods to prevent fatal outcome in road traffic

collision are taken (e.g. helmets or seatbelts). Post-crash care is another crucial factor in minimizing mortality and morbidity in traffic crashes, but highly dependent on available resources which explains the discrepancy between low- and high-income countries in providing prehospital care. The evidence-based guideline *Save LIVES: a road safety technical package* has been provided by the WHO for decision-makers to reduce road traffic deaths [3]. The acronym LIVES sets the focus on the interventions speed management, *L*eadership, *I*nfrastructure design, *V*ehicle safety, *E*nforcement of traffic laws and post-crash *S*urvival.

*References*

*[1] Peden M, Scurfield R, Sleet D, Mohan D, Hyder AA, Jarawan E, et al.: World report on road traffic injury prevention. Geneva 2004.*

*[2] World Health Organization (WHO): Global status report on road safety 2018. Geneva 2018.*

*[3] World Health Organization (WHO): Save LIVES: A road safety technical package. Geneva: World Health Organization Copyright 2017.*

### Rationale

In the context of globally rising motorization of transport and the SDG target to decrease road traffic deaths to 50 % by 2020, further research should be performed to find factors with enormous potential impact. This project aims to re-evaluate and identify factors leading to a high severity in collisions with serious and/or fatal injuries, and to describe the prediction strength. A prediction model on collision severity is also needed to efficiently distribute the limited resources of emergency health care to the places where needed the most.

Such a model enables decision-makers

- to focus on most efficient preventive strategies
- to allocate limited health care resources
- to provide a risk assessment for individual considerations as a traffic participant

By doing so a range of stakeholders such as politicians, emergency health care managers, physicians and individual representatives of traffic participants should be addressed in the evaluation process.

As there already exists evidence and good research regarding factors contributing to the probability of a traffic collision and different measures are taken to reduce the risk (see Save LIVES), this project focuses on the aspect of distributing the limited resources of emergency health care to the right places thus concentrating on severity-related factors which are observed without in-depth investigation and available as information for the health care decision-makers.

## Data Understanding

### Data source

The data used to build this model is provided by the Seattle Department of Transportation which is the official governmental department focusing on transportation systems and providing safe access to places in Seattle. The dataset consists of all types of collisions from 2004 on combined with several attributes such as weather, road or light condition, involved traffic participants, infrastructural aspects and geographical data. Some of the attributes have already been linked to the occurrence of traffic collisions (e.g. speeding, weather, road or light condition) and will now be used to evaluate the collision severity. The latter is labelled by a severity code ranging from 1 (property damage only) to 3 (fatality) in the provided dataset. Attributes of interest to predict severity in the first place without further in-depth assessment include the following points while individual factors which need further investigation such as inattention or drink-driving and the specific location were excluded.

- involvement of vulnerable traffic participants: pedestrians, bicycles
- infrastructural factors: junction type, road condition, pedestrian right of way, speeding
- environmental factors: weather condition, light condition

Being able to include the contextual data of a traffic collision into the model will ensure a comprehensive assessment of potential influencing factors. The choice of factors may enable conclusions for a more general setting without time or location specific data.

## Data pre-processing

The target variable – the severity of traffic collisions (SEVERITYCODE) – was assessed and due to the limitation of its dimensions to 1 ('property damage only') and 2 ('injury') defined as a binary variable (Y_SEVERITYCODE).

As one of the first steps the dataset was reviewed with regard to its integrity. Missing values were obtained by exploring the respective columns, marking 'Unknown', '0' or any other possible missing value to NaN. Attributes with missing rates of more than or equal to 5 % were looked at in more detail. The affected variables (missing rate in brackets) included

- INTKEY (66.58 %)
- EXCEPTRSNCODE (56.43 %)
- EXCEPTRSNDESC (97.10 %)
- INATTENTIONIND (84.69 %)
- PEDROWNOTGRNT (97.60 %)
- SDOTCOLNUM (40.96 %)
- SPEEDING (95.23 %)
- WEATHER (10.36 %)
- ROADCOND (10.32 %)
- LIGHTCOND (9.58 %)

To get an idea of the missing mechanism a graphical approach via heatmaps and dendrogram was applied which revealed correlations between the missing rate of linked attributes (see Fig. 2)

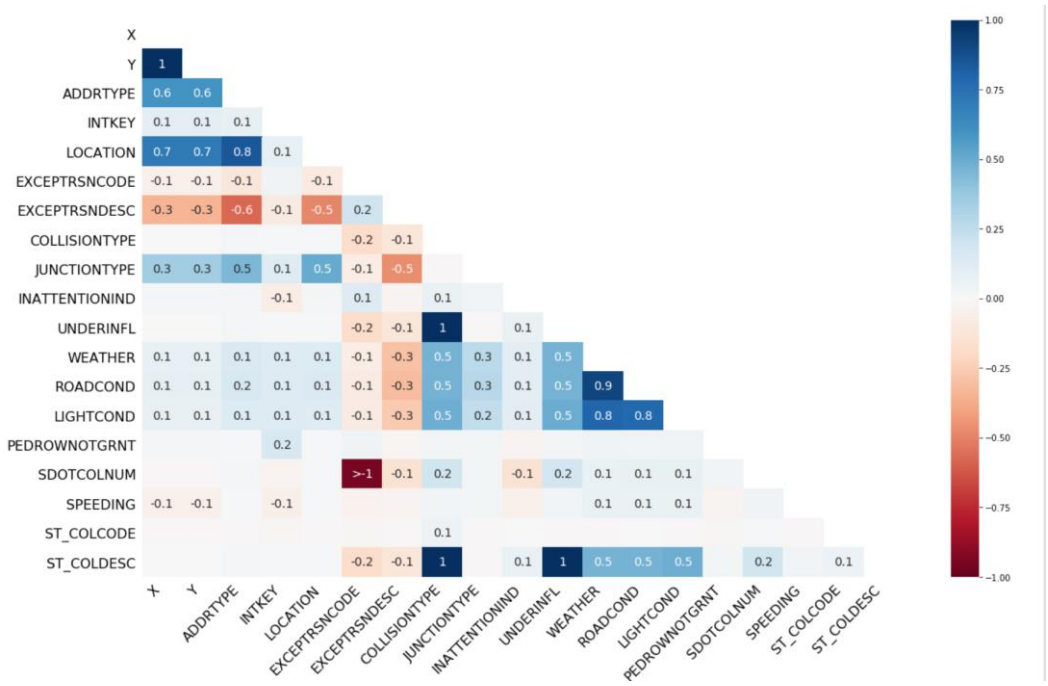


*Fig. 2: Correlation heatmap of missing values*

When grouping the missing rates by the target variable no significant difference between these populations could be detected via chi2-testing. Taking these considerations into account there seems not be a missing mechanism completely at random (MCAR) but rather missing at random (MAR). These assumptions need to be considered in case of further pre-processing and handling missing data.

The original dataset consisted of 38 columns, during feature selection the columns were reduced to a relevant number of 9. The feature selection involved several steps: Identifier parameters such as chosen IDs or report numbers were removed, doubly-present (redundant) variables were removed, some variables of interest were aggregated and updated, and time- or location-specific data was excluded from the analysis. In the context of the project's aim to focus on the health care provision the following attributes available at the moment of prediction (without further in-depth investigation) remained:

- PERSONCOUNT: amount of involved persons
- VEHCOUNT: amount of involved vehicles
- WEATHERCOND_1: weather condition
- ROADCOND_1: road condition
- LIGHTCOND_1: light condition
- JUNCTION: relation to a junction
- PED: involved pedestrians
- BYC: involved bicycle

## Balancing

The chosen dataset presents as a real-world dataset and therefore reveals by being imbalanced a common data science problem (see Fig. 3). While the majority of the 194,673 reported collision are property damage only (70.11 %). Many models do struggle with identifying the minority classes thus a pre-processing step to overcome this challenge must be implemented. A possible approach is to simply undersample the majority class which is in the context of the large dataset an adequate option but still losing statistical power. Another method is directly integrated in the modelling algorithm by SciKit Learn's where a weight can be added to the minority category. The latter was chosen due to its good performance and preventing undersampling biases.

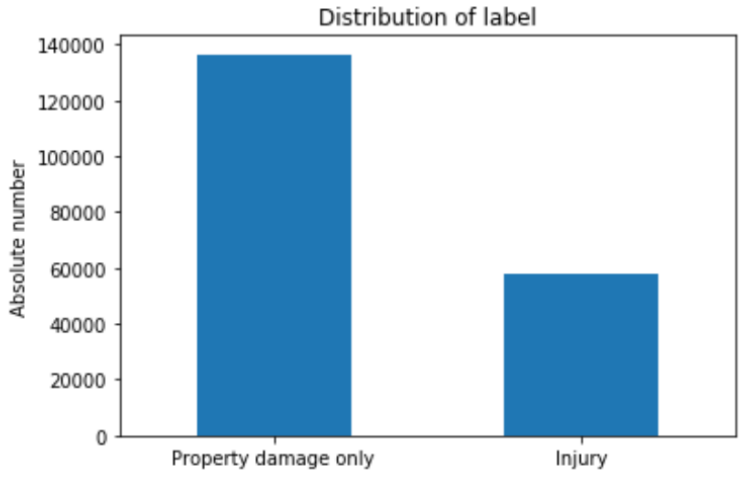


*Fig. 3: Unbalanced dataset*

## Exploratory data analysis

When having a closer look to the numerical values a mean involvement of 2.44 persons and 1.92 vehicles could be obtained. Most of the collisions lead to property damage only, did not involve pedestrians nor bicycles, and did not take place in relation to a junction. The most often circumstances involved clear weather conditions, a dry road and daylight. When grouping the categorical variables by the collision severity a group difference could be detected via chi2-testing in all of the attributes. Via Mann-Whitney-U-Testing a group difference regarding the number of involved persons and vehicles could be detected.

Bar graphs illustrated the severity distribution among the respective attributes (see Fig. 4 – Fig. 7).

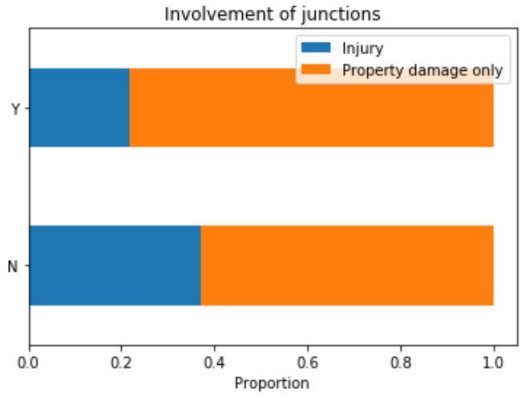


*Fig. 4: Traffic collision severity related to junctions. Y = Yes, involved junction, N = No junction.*

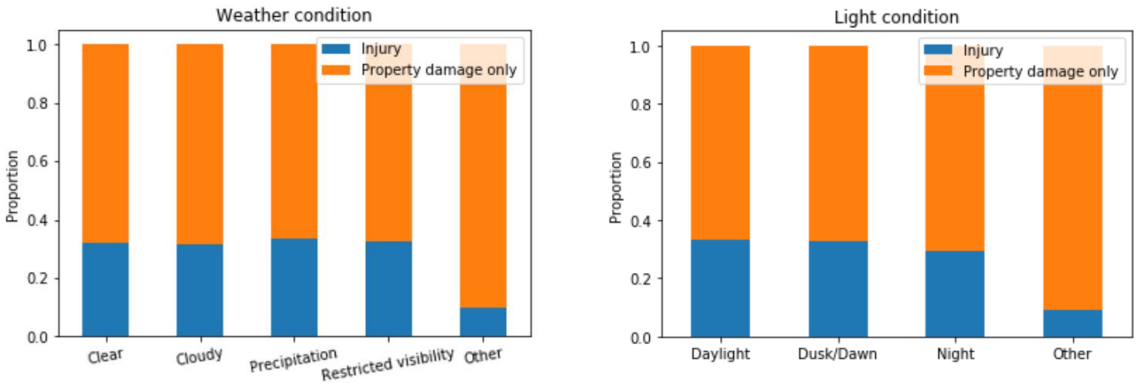


*Fig. 5: Weather and light condition.*

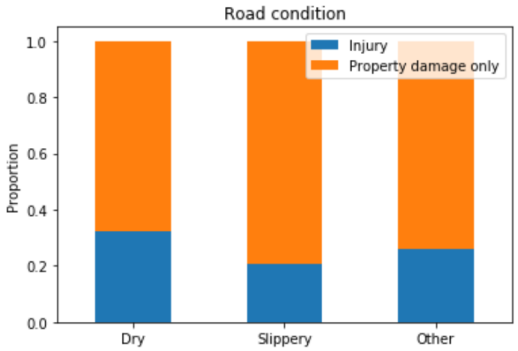


*Fig. 6: Road condition.*

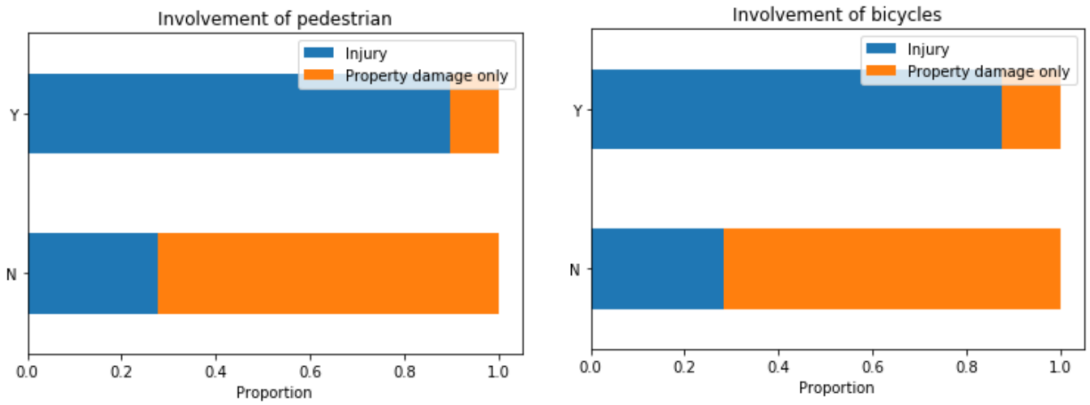


*Fig. 7: Traffic collision severity related to involved pedestrians or bicycles. Y = Yes, involved, N = No, not invovled.*

With these graphical illustration in mind, the clear impact of involved pedestrians or bicycles on the collision severity was outstanding.

## Results

To find attributes to predict traffic collision severity there are several approaches for modelling. As regression provides additional information on the impact of the respective factor, it was the method of choice to enable decision-makers not only to predict the severity but also to take decisions regarding most effective prevention measures. All of the chosen attributes showed a significant group difference regarding traffic collision severity. For further predicting the severity thus enabling stakeholders to categorize the collision and distribute resources, a regression model was built.

### Modelling

When taking categorical variables as potential influencing factors for predicting these need to be turned into categorical variables (‘dummys’). As a second step the Skikit learn algorithm for logistic regression was used. Two different approaches were chosen, one to balance the dataset (LR), the other without balancing (LR1). As a solver ‘sag’ was chosen to sufficiently address multiclass problems and the large dataset. The regularization strength was set to 0.01, and the dataset was split into training and testing by 0.2.

## Evaluation

The following evaluation metrics (see Tab. 1) were obtained for the respective model. As assumed the balanced dataset performed better predicting true positive (true injury) values while losing precision. As it is crucial to correctly predict the collision with injury while it is acceptable to include a larger number of false positive injury-labelled collisions (see Fig. 8), the balanced model was chosen as the final model to predict traffic collision severity.

| Model | Log loss | Jaccard similarity score | Precision for minority category | Recall for minority category | F1-Score for minority category |
|---|---|---|---|---|---|
| **Balanced** | 0.60 | 0.65 | 0.43 | 0.65 | 0.52 |
| **Unbalanced** | 0.53 | 0.75 | 0.81 | 0.22 | 0.34 |

*Tab. 1: Evaluation metrics for balanced and unbalanced regression model.*

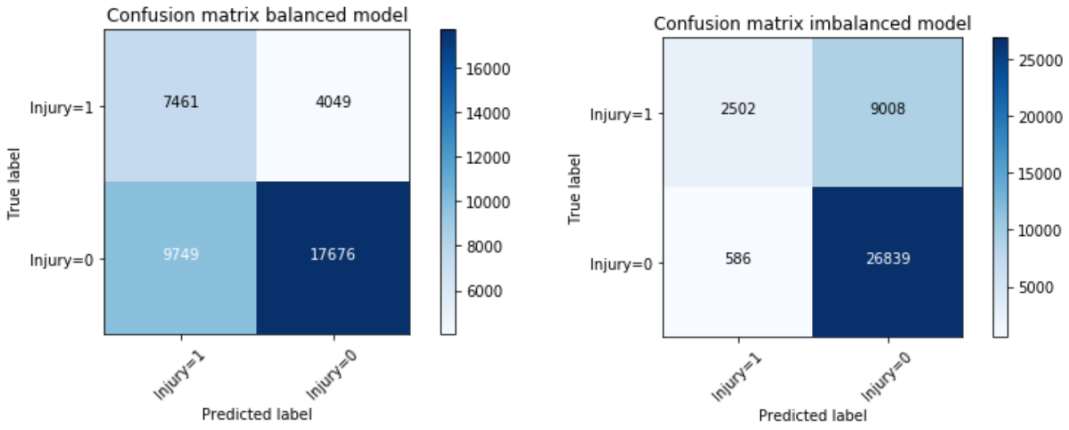


*Fig. 8: Confusion matrix for balanced and unbalanced regression model.*

## Conclusions

As outlined in the introduction, the prediction of traffic collision severity is crucial to successfully distribute limited health care resources. The developed model focuses on predicting the severity (Property damage only versus injury) using attributes which can be obtained through first information. Via a logistic regression algorithm, the severity can be predicted and the final model ensures that the number of incorrectly labelled collisions as 'property damage only' are minimized while it accepts a higher number of incorrectly more severe labelled collisions.

Furthermore, the attributes chosen are all relevant factors of traffic collision severity, thus being targets of further prevention strategies.

Such a model needs to be revised thoroughly as it may lead to important decisions taken by health care managers. The presented model is a draft which definitely should not be used for far-reaching decisions which affect the health provision in emergency settings. It needs to be finalized and iteratively evaluated.