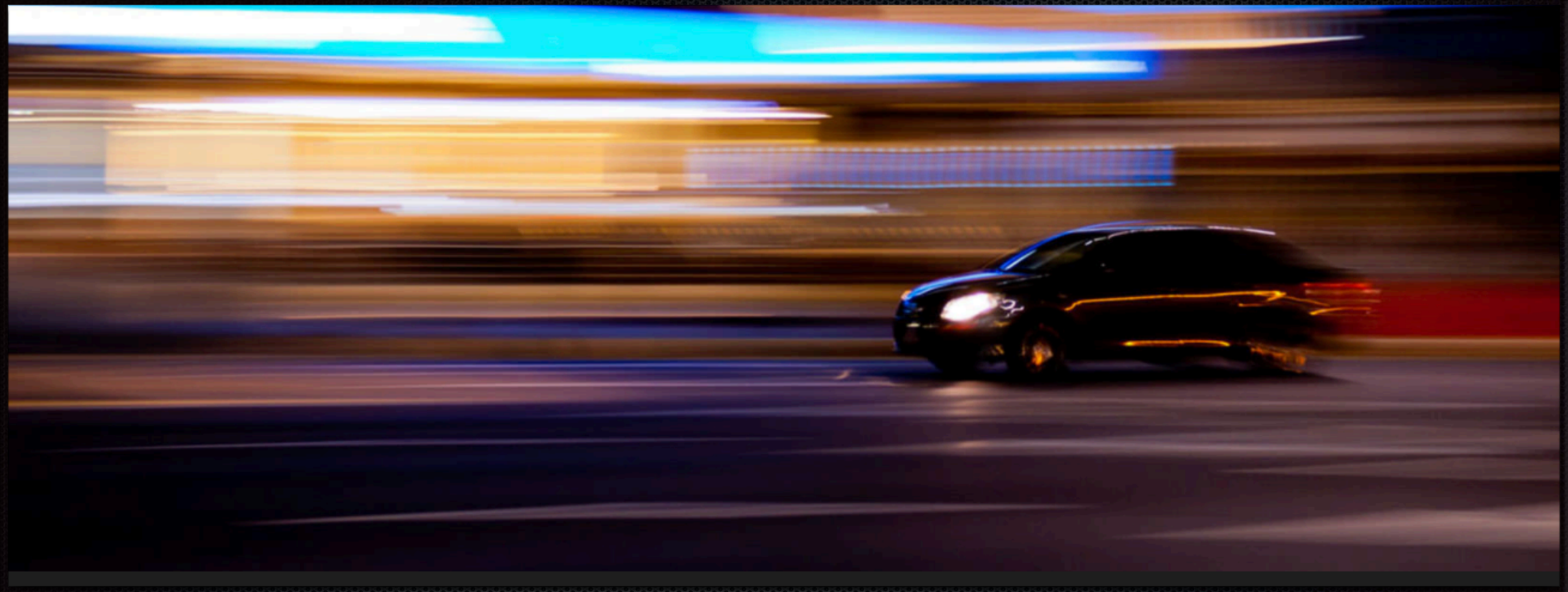


Car Accident Severity Prediction

Riccardo Bellio

Predictive model creation for car accident severity

On Feb 2020, the World Health Organisation revealed that worldwide:

- **1.35 million** people die as a result of a car accident.
- **20** to **50 million** people suffer non-fatal injuries.

To tackle this problem, relevant data was analysed to explore factors that could impact people's safety while driving and to create a predictive model using classification algorithms.

The project could be used to:

- Share the collected information with the general public to increase their awareness of car accident risks factors.
- Present the model to governments & private companies to further improve their content, develop an alert system to inform drivers on their risk level of being involved in a car collision and promote corrective behaviour (such as driving more carefully or change route).

Data Acquisition & Data Cleaning

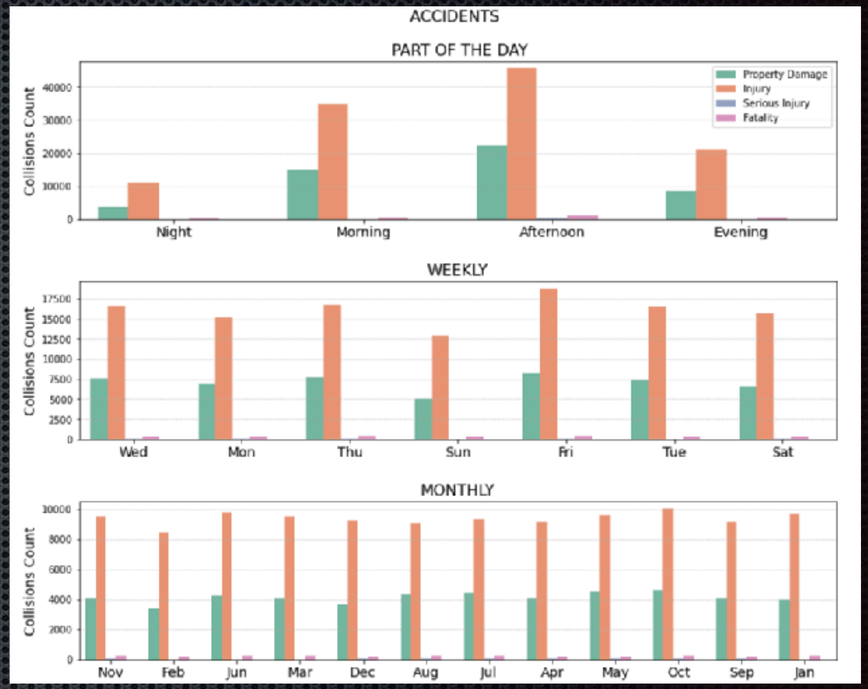
The dataset used in this project was provided by **IBM** and is available on the open data portal of the city of Seattle, U.S.A. It contains detailed information relating to vehicle collision severity recorded on a weekly basis from 2004 until today in **Seattle**, with a total of 221,265 entries and 40 columns.

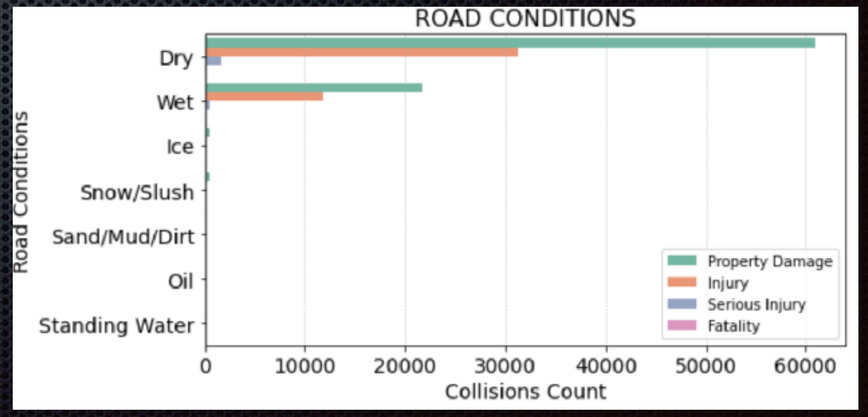
- Feature selection reduced the total number of attributes from **40** to **13**, based on their importance for modelling.
- Time, day of the week, month and year of recorded accidents were extracted and displayed in separate columns.
- '**Nan values**' were analysed column by column and dropped if considered non relevant or kept if significant.
- '**Others**' and '**unknown**' values were identified and dropped.
- The '**Unknown**' class of label data was dropped.
- Original value scale of label data was changed for clarity.

OLD VALUE	NEW VALUE	DESCRIPTION
1	1	Property Damage Only
2	2	Injury
2b	3	Serious Injury
3	4	Fatality

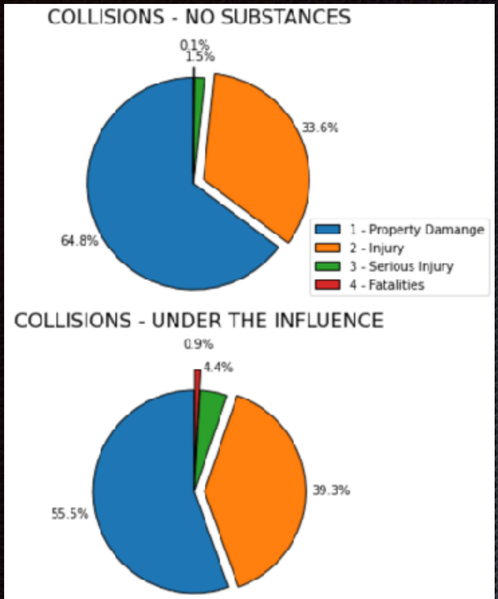
Exploratory Data Analysis

- **Yearly collisions** showed a downtrend from 2006 to 2010, followed by an uptrend and then a downtrend from 2015 to 2019. Data for 2020 was incomplete since the year was not over yet.
- **Highest collision rates** occurred afternoons, on Fridays and in October, November, January and June.
- **Highest collision rate** was recorded at intersections and when a parked car was involved.
- **Injuries** were mainly the product of collisions at rear ends and angles. **Fatalities** peaked when accidents involved pedestrians.
- **Clear** weather conditions showed the highest rate of collisions (64.9%). **Raining** and **overcasted** conditions together represented only 34.3% of the total.
- The majority of the accidents took place when the road condition was **dry**, followed by **wet** conditions.

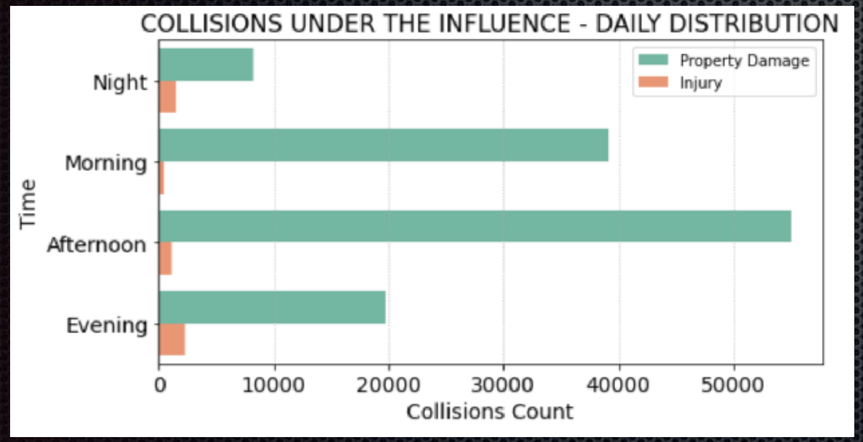




Exploratory Data Analysis



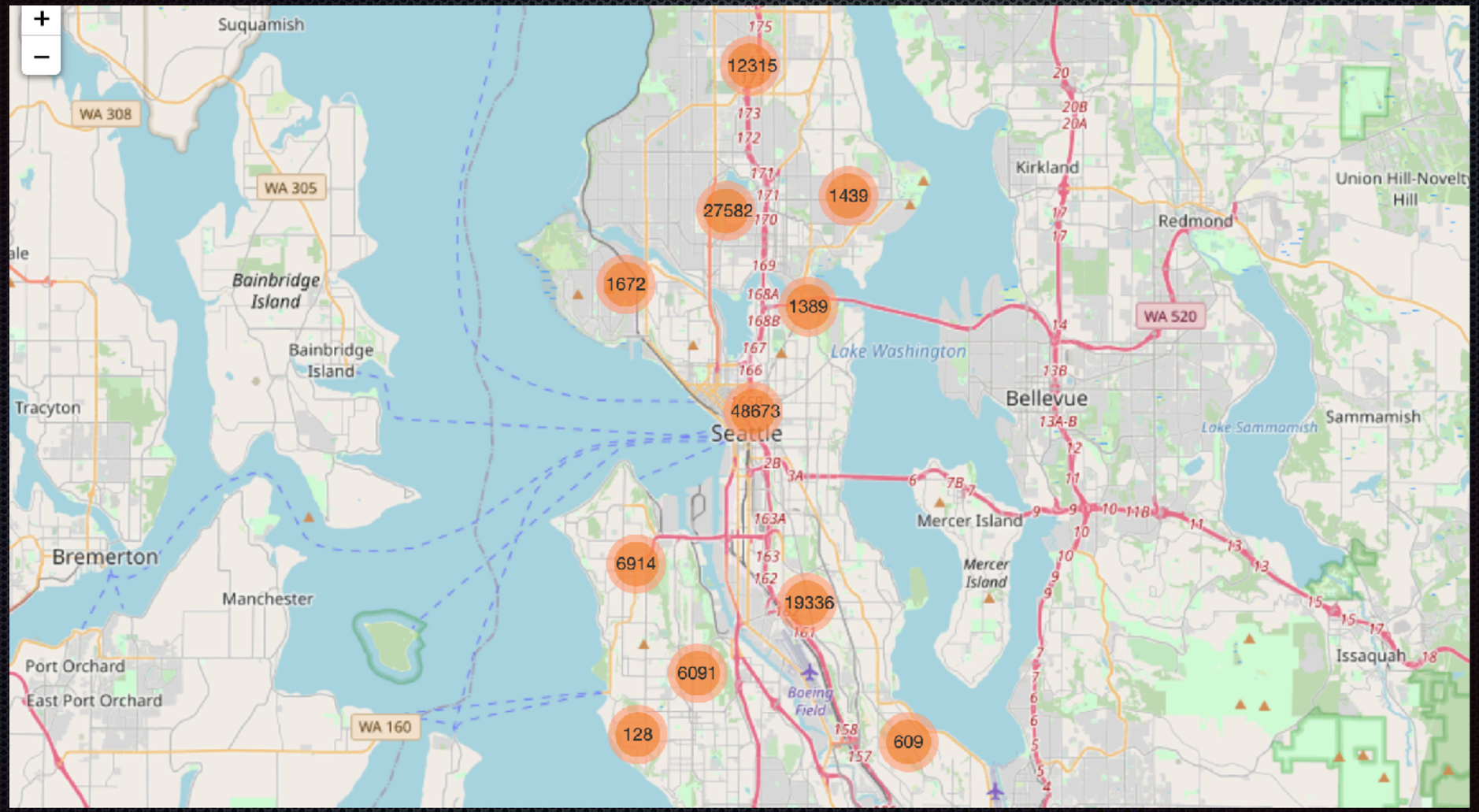
- **Daylight** accidents were the most frequent. The **absence of light** did not affect the accident severity distribution.
- Collisions caused by drivers driving under the influence of alcohol or other substances were less frequent, but **more severe**: fatalities +0.8%, serious injuries +2.8% and injuries +6.1%.
- **Speeding** increased collision severity: fatalities +0.6%, serious injuries +1.9% and injuries + 7.2%.



- The majority of the accidents involved less than five people (the highest rate involved two people) and only drivers (followed by one pedestrian and one cyclist only).
- The correlation of independent variables was inspected and did not show any significant correlations between features.

Accident Interactive Map

An interactive map to explore the street of the city of Seattle and visualise the distribution of the collisions was plotted. It showed that the majority of the accidents occurred in the **city centre**.



Dealing with an extremely Imbalanced Dataset

- Object and categorical data types were converted into int64 data types using the method **pd.get_dummies** for one hot encoding.
- The cleaned dataset (with a total of 127.505 entries) was **extremely imbalanced** since the vast majority of vehicle accidents involved property damage only with fewer fatalities.
- Classification algorithms tend to over-fit on the most represented class label and **not effectively predict** the minority classes. To avoid this outcome, the dataset was **balanced out**: the majority classes were under-sampled and the minority class (fatality) over-sampled. A slight imbalance was maintained to reflect the original dataset distribution.
- Traditional classification algorithms do not perform well on multiple-class classification problems with balanced datasets. To improve their performance the **labeled data was binarised** and the four classes in the labeled output were turned into two classes: property damage, injury and serious injury were merged into a new class non-fatality, thus creating a labeled data divided into fatality and non_fatality entries.

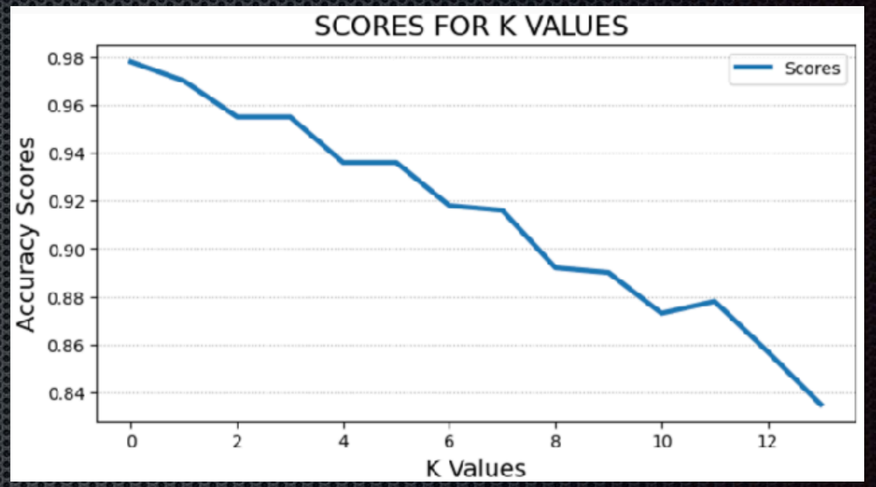
CLASS	ACTION
1. Property Damage	Randomly under-sampled to 35000
2. Injury	Randomly under-sampled to 30000
3. Serious Injury	Over-sampled to 25000
4. Fatality	Over-sampled to 25000

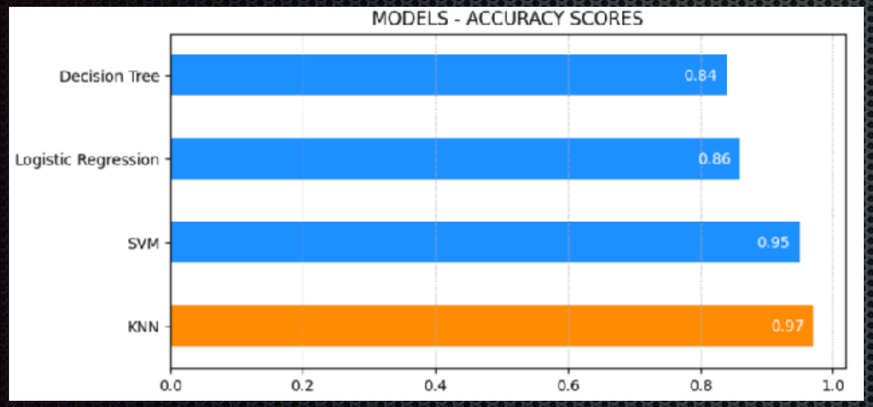
OLD CLASS	NEW CLASS
1. Property Damage	1. Non_Fatality
2. Injury	1. Non_Fatality
3. Serious Injury	1. Non_Fatality
4. Fatality	2. Fatality

Classification Model & Performances

Four different algorithms were used for this classification problem: K-Nearest Neighbours, Decision Trees, Logistic Regressions and Support Vector Machine. The process followed for model creation was the same for each algorithm:

- Data was **scaled** and put in a format that worked better for algorithms (each feature value close to zero with all features on comparable scales).
- Dataset was split into **Train** (67% of the total) & **Test** (33 %) **sets**.
- The specific algorithms **hyper-parameters** were **optimised** (such as k values for KNN an Max_leaf_nodes for the Decision Tree).





Model performances were classified using the measurements of their accuracy, precision, recall and confusion-matrix.

The four different classification models performed extremely well. The high model accuracies were confirmed by their confusion matrix, thus eliminating the suspect that the model was not equally predicting true values. **KNN** was the model with the best performances in terms of accuracy, precision, recall and F1-score.

This positive result was probably due to the fact that the under-represented data was **oversampled 10 times**. This meant that the dataset did not posses a great variety and that part of the test set already contained most of the data included in the test set.

Conclusions & Future Directions

The relationship between accident severity and data relating to the recorded accidents was analysed. Data exploration revealed some interesting information that should be shared with the general population to increase awareness on the risks.

The method used to collect the feature '**inattention**' deserves further investigation; I expect drivers to be reluctant to admit their inattention. For example, if they were distracted because they were using their mobile phone, insurance companies would not covered costs. I am curious about how this data was collected.

Further research would have to be implemented with **more relevant information**. For example, most collisions took place when the weather condition was clear. Was this due to the fact that drivers were more careful when driving in bad weather or because the weather in Seattle is dry mostly all year around? Data in relation to weather conditions would have to be included in order to give meaning to this outcome.

Imbalanced data was **balanced out** and multi-class label was **binarised**. A different way of dealing with a classification problem with imbalanced data could have been to maintain the original four classes and use more complex predictive models. In fact, the models used in this project were tested on the cleaned multi-class dataset and never returned an accuracy score higher than 50%. This type of solution could be implemented in a future project.

The predictions of four different classification models were compared and the best one was chosen. The best model accuracy was very high: **97% achieved using KNN algorithm**. The prepared dataset did not posses a great variety because of the balancing process on the original dataset. As a result, part of the test set was already contained in the data included in the test set. A future project could include different datasets relating to this topic recorded in different parts of the world, and they could be tested in different ways to evaluate the presence of common trends and points of differentiation.

“Thank you for reading.”