How to predict accident severity?

IBM Course: Applied Data Science Capstone

Alessandro Sgarabotto

17/10/2020

Business understanding

- Traffic accidents are one of the major cause of mortality and disability (top 10 causes of death WHO);
- Traffic accidents results in not only mortality and disability but also in expense increase for the public healthcare system
- Predicting the severity of an accident will help road safety reducing the death casualties and their economic impact
- Numerous factors affect the severity of an accident.

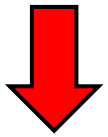
Can we predict on scientific basis the severity of accident?

Data Sources

- The data is provided as course material. The dataset refers to severity of accidents occurred in the city of Seattle (US) from 2004 to present time. All collisions were provided by Seattle Police Department (SPD) and recorded by Traffic Records.
- The dataset has 194 673 observations and 38 features containing both numerical and categorical data.
- The first column is SEVERITYCODE and pertains to the severity code, namely the accident severity which is labeled as 3, in case of fatality, 2b, in case of serious injury, 1, in case of damage or 0, in case of unknown details (in the present dataset the severity code is a binary variable that is labeled as 1 for low severe accident or 2 for high severe accident).
- The objective of this project is the prediction of the severity code (i.e., the accident severity) by means of supervised machine learning algorithms.

Data Cleaning

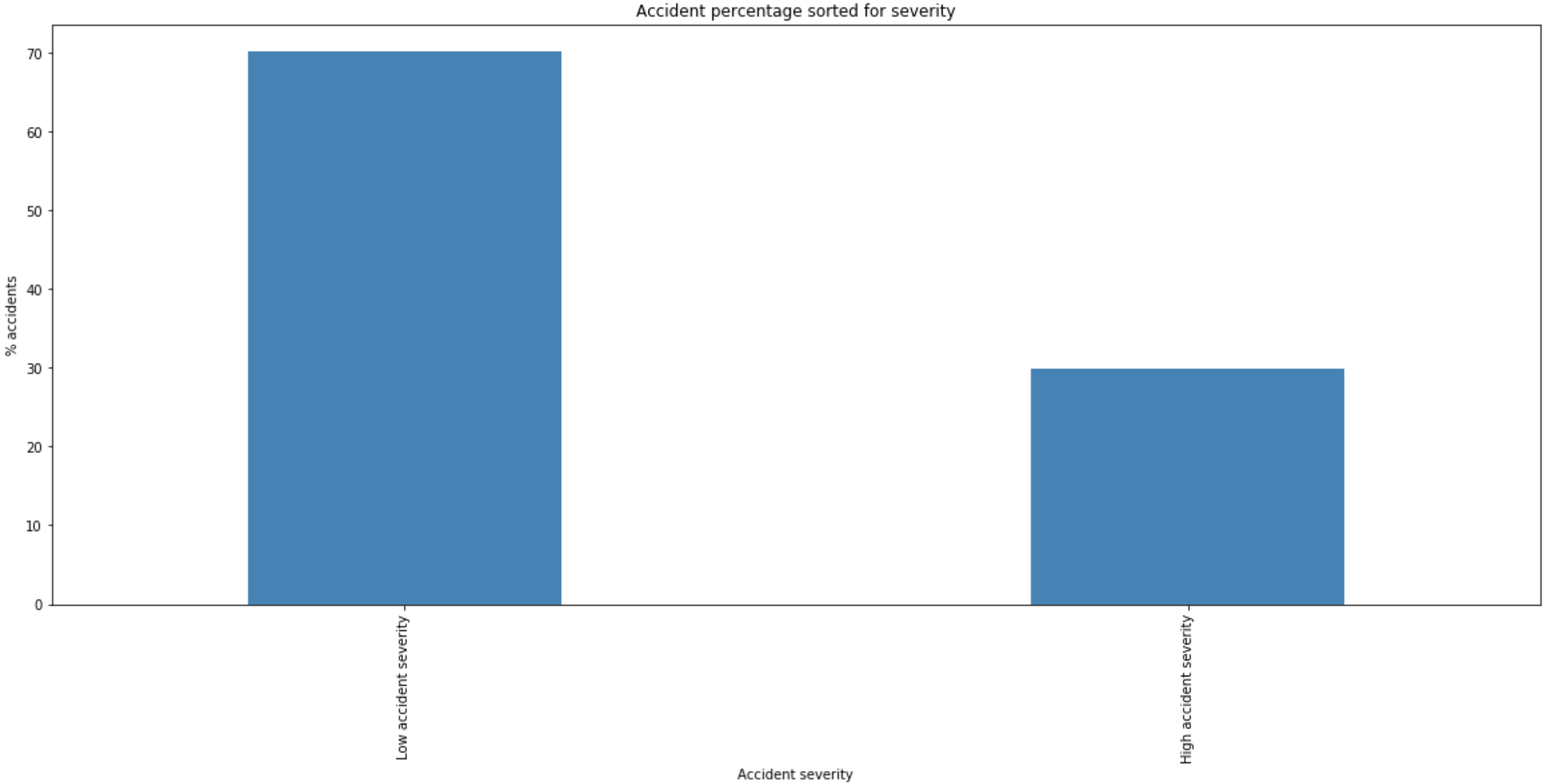
- Not all the features are essential to the present investigation
- There are missing data in the features. The missing values are substituted with the mean for numerical values, and with the most common values for categorical values
- There is some redundancy with some features. The redundant features are dropped.
- Categorical variables are turned into numerical variables.



Independent variables (target): SEVERITYCODE
Dependent variables (features): ADDRTYPE, ROADCOND, LIGHTCOND, SPEEDING, UNDERINFL, INATTENTIONIND, PERSONCOUNT and VEHCOUNT

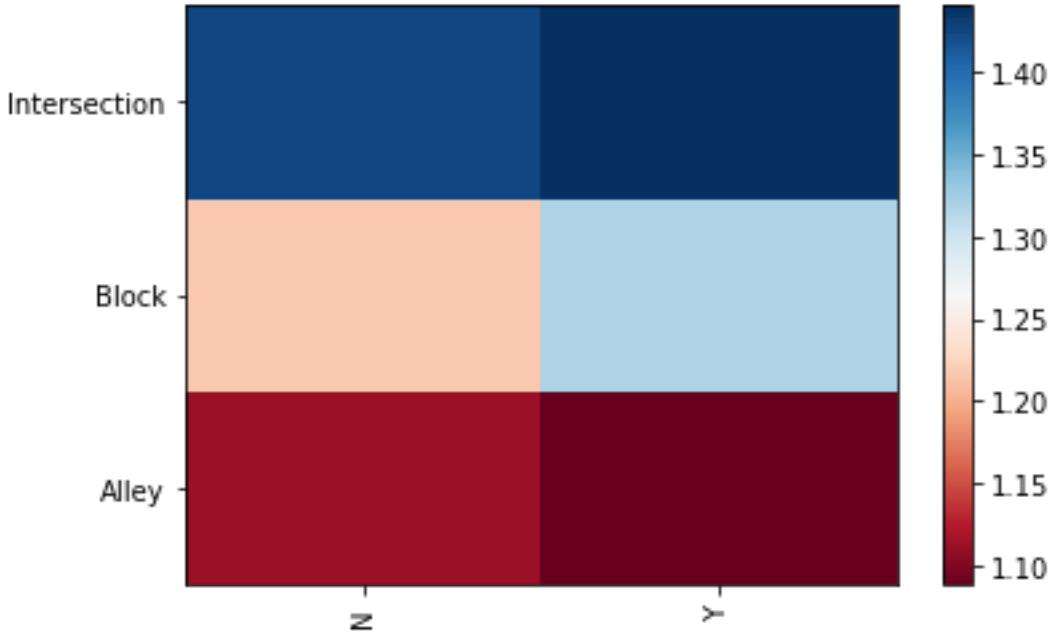
Data exploration

The majority of accidents has low level of severity, namely damage with no injury

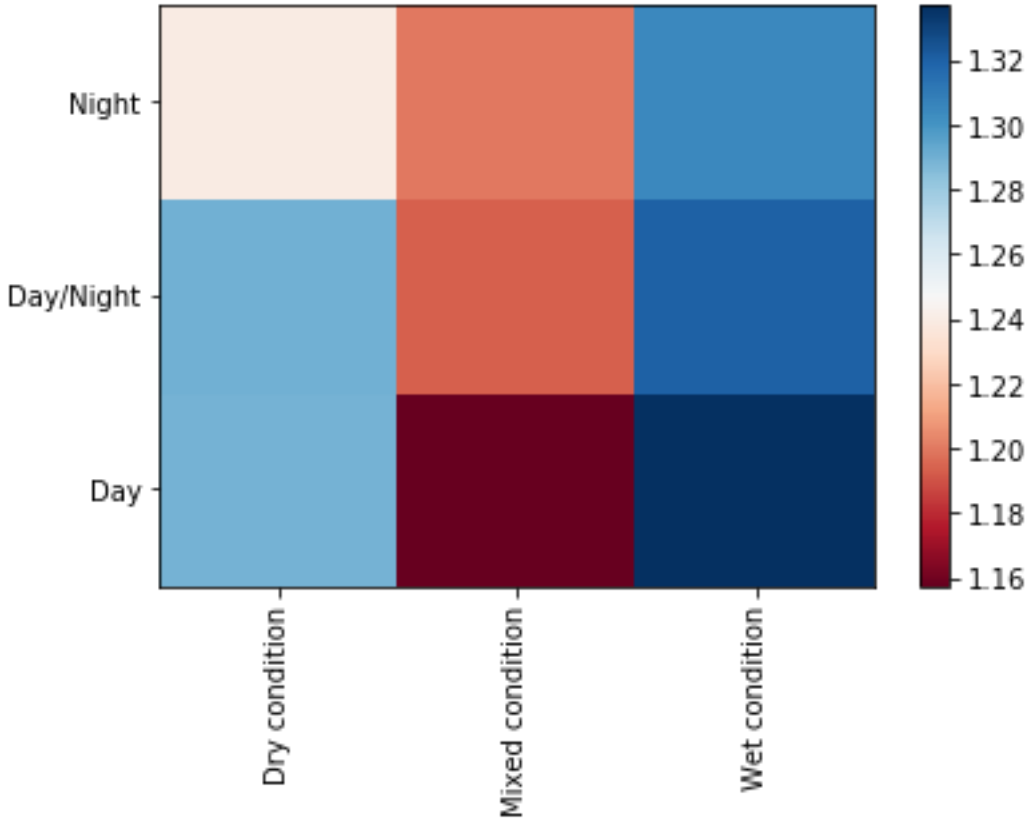


Data exploration: heat maps (I)

The relationship between the accident severity and other variables is plotted in heat maps which show the mean value of the accident severityproportional to color with respect to the other variables in the vertical and horizontal axis.



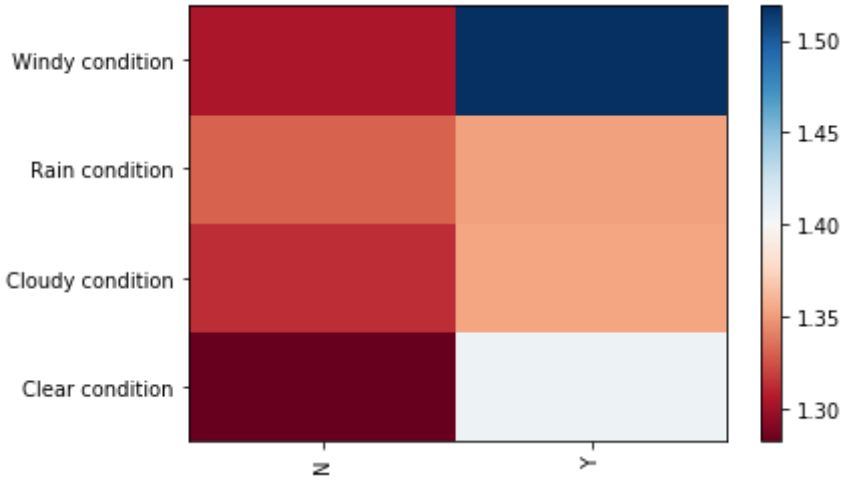
Relationship between mean accident severity changes, collision type (ADDRTYPE) and inattention at driving (INATTENTIONIND)



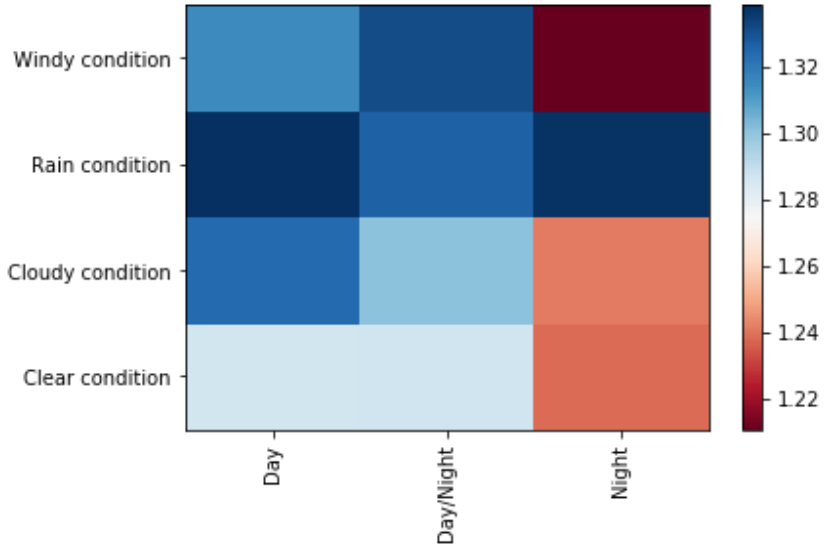
Relationship between mean accident severity changes, lighting condition (LIGHTCOND) and road condition (ROADCOND)

Data exploration: heat maps (II)

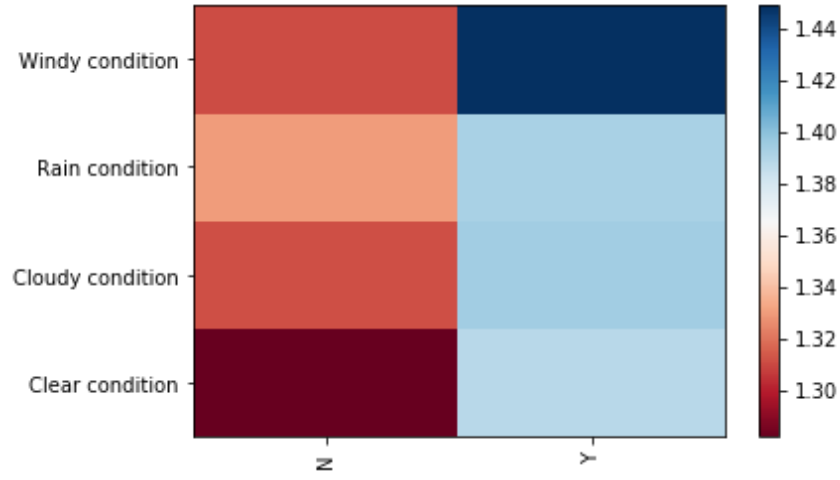
It is worth to see the relationship between the mean accident severity, weather condition and other variables . The rain and wind seem to increase the accident severity at daylight, in condition of high or medium visibility.



Relationship between mean accident severity changes, weather condition (WEATHER) and car speeding (SPEEDING



Relationship between mean accident severity changes, weather condition (WEATHER) and lighting condition (LIGHTCOND)



Relationship between mean accident severity changes, weather condition (WEATHER) and influence of drug/alcohol (UNDERINFL).

Data exploration: Which feature has the highest impact on accident severity?

	Pearson	P-value
PERSONCOUNT	4047.862	$8.177 \cdot 10^{-129}$
VEHCOUNT	0.131	0.0

	F-test	P-value
ADDRTYPE	4047.862	0.000
INATTENTIONIND	419.618	$3.706 \cdot 10^{-93}$
LIGHTCOND	3.964	0.119
ROADCOND	175.484	$7.188 \cdot 10^{-77}$
WEATHER	108.369	$4.179 \cdot 10^{-70}$
SPEEDING	295.607	$3.340 \cdot 10^{-66}$
UNDERINFL	384.119	$1.908 \cdot 10^{-85}$

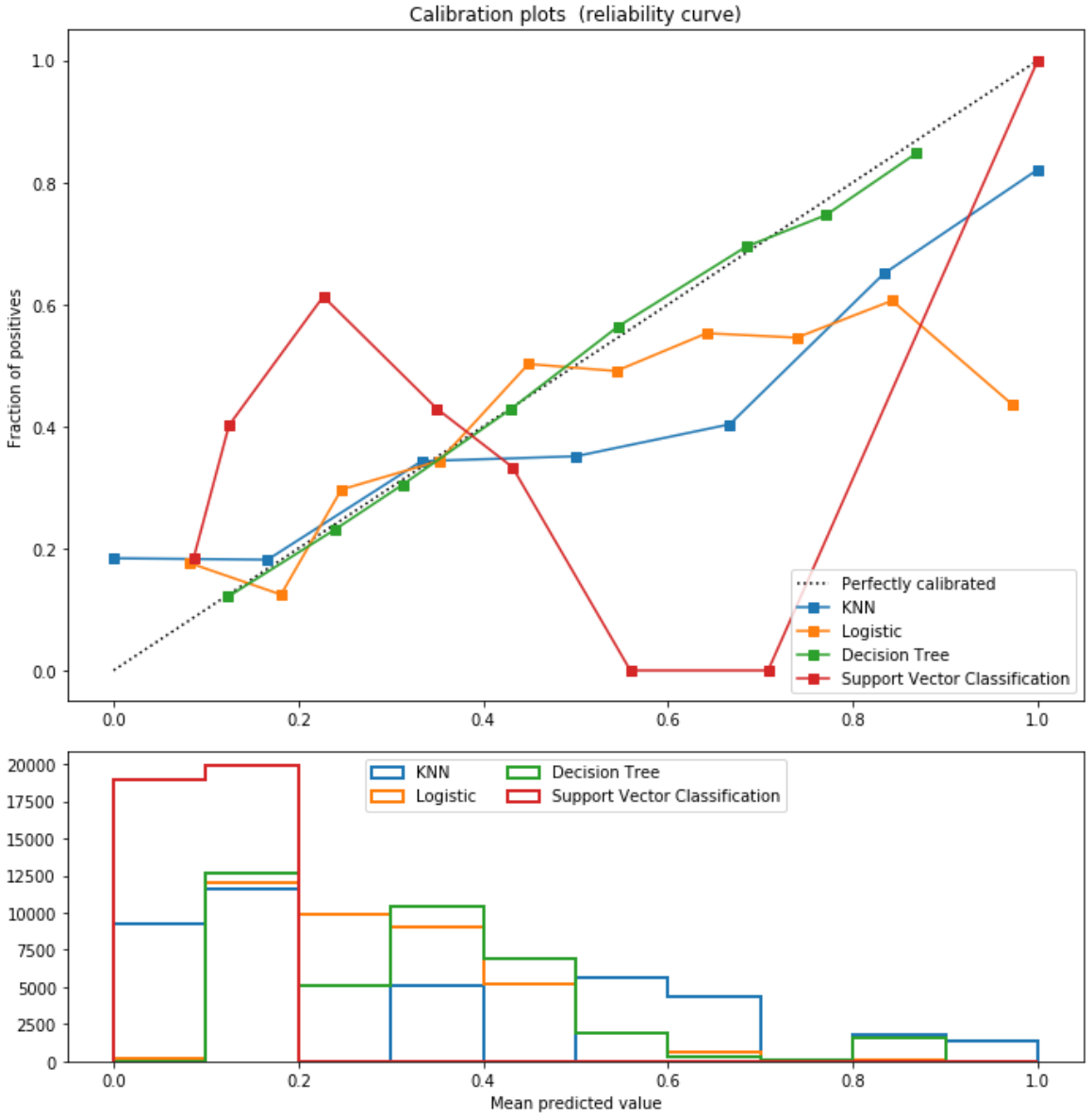
- Each features has a significant correlation with the accident severity (low P-value);
- The more severe is the accident, the higher is the number of people (PERSONCOUNT)and vehicles involved in the accident (VEHCOUNT);
- The features have a strong impact on the accident severity (high F-test score) with the exception of lighting condition (low F-test score);
- The accident severity is affected the most by the following three variables: collision type (ADDRTYPE), inattention at driving (INATTENTIONIND) and influence of drug/alcohol (UNDERINFL).

Building up the predicting model

- Features are subjected to standardization
- Features and target are randomly subdivided in two subsets: one for training the data (80%) and the other for testing the data (20%)
- The labeled data accident severity is used to train and test a model by supervised machine learning. The machine learning algorithms considered are:

- K-means Nearest Neighbours (KNN)
- Decision Tree (DT)
- Logistic regression (LR)
- Support Vector Machine (SVM)

Calibration



All classifier seem to well predict the accident severity in the test subset except for the support vector machine that could not reach convergence showing poor fitting. However, given the model results, the accident severity was anyway well predicted by the other models.

	Jaccard	F1-score	Log-loss
KNN	0.70	0.74	N/A
DT	0.74	0.69	N/A
LR	0.71	0.63	0.57
SVM	0.71	0.63	N/A

Conclusion

- This study aims at predicting the severity of an accident. The data source is provided by the course and refers to accidents record occurred in Seattle (US) from 2004 onwards.
- The exploratory analysis reveals that the features that have the most impact on the accident severity are: collision type (ADDRTYPE), inattention at driving (INATTENTIONIND) and influence of drug/alcohol (UNDERINFL).
- Since the data provided is labeled, supervised machine learning algorithms represents the best option for the investigation purpose. Different classifier were applied, namely K-Nearest Neighbors, Decision Trees, Logistic Regression and Support Vector Machine. The dataset was subdivided randomly in train subset, to fit the model, and a test subset, to evaluate the model performance. All classifier seem to well predict the accident severity in the test subset except for the support vector machine that could not reach convergence showing poor fitting. However, given the model results, the accident severity was anyway well predicted by the other models.
- For future developments, other supervised machine learning algorithms can be taken up for this investigation. On the other hand further data can be studied to see results obtained for Seattle are universal. Surely, the results could be of interest for car factories which should should be encouraged to design hi-tech cars able to prevent and correct lack of attention at driving.