Predicting the severity of an accident

Alessandro Sgarabotto

17/10/2020

1 Introduction: Business problem

Before boarding a plane, a high school student has a premonition: the plane is going to crash as soon as it takes off. As the boarding proceeds, this gut feeling grows stronger so that he persuades his friends to stay in the airport waiting for the next flight. While he is being blamed for what seemed to be a panic attack, the plane crashes as it takes off. This is the beginning of the cult movie *Final destination* (2000) which, going beyond its sci-fi plot, introduces the topic of this report: is it possible to predict on scientific basis the severity of an accident ?

(a)

(b)

(c)

Figure 1: a) Traffic jam in Milan (Italy, link a), b) Crash involving a motocycle in Milan (Italy, link b), c) Movie scene in *Final destination* (2000) where the main actor has a premonition of an accident and tries to persuade his friends to change their trip plans (link c).

1.1 Background

Traffic accidents are one of the major cause of mortality and disability and, according to World Health Organization, rank among the top 10 causes of deaths in the world (for details see WHO site). Traffic accidents results not only in mortality and disability but also in hindrance for public viability, since people take long to reach their destination, and for public healthcare, since people may be hurt in the accident. In a nutshell, accidents turn up to cause an expense increase to cope with injured people and traffic jam. Therefore, traffic accidents have a wide spread of economic impacts that should be addressed.

1.2 Problem

Numerous factors may affect the severity of an accident, such as whether conditions, road layout, lighting conditions, vehicle speeding, drug/alcohol influence and so on. Understanding how these features affect the accident severity is not a problem that has a straightforward answer. The aim of this work is to predict the severity of an accident by such kind of features.

1.3 Interest

Understanding how the accident severity is related to traffic or drivers features is of utmost importance. If we predicted the severity of an accident, preventive measures could be be taken for people safety controlling the accident severity and thus its economic consequences.

2 Data acquisition and cleaning

2.1 Data sources

The data is provided as course material at this link. 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 (Table 1).

The first column is SEVERITYCODE and pertains to the severity code, namely the accident severity which is labeled as 3, in case of fatality, 2*b*, 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

Feature	*Description*
SEVERITYCODE	A code that corresponds to the severity of the collision
X, Y	Latitude and longitude of the accident location
OBJECTID	Unique identifier
INCKEY	Unique key for the incident
COLDETKEY	Secondary key for the incident
REPORTNO	Report number
STATUS	Report status
ADDRTYPE	General type of collision
INTKEY	Key that corresponds to the intersection associated with a collision
LOCATION	Description of the general location of the collision
EXCEPTRSNCODE	
EXCEPTRSNDESC	
SEVERITYCODE.1	A detailed description of the severity of the collision
SEVERITYDESC	A detailed description of the severity of the collision
COLLISIONTYPE	Collision type
PERSONCOUNT	The total number of people involved in the collision
PEDCOUNT	The number of pedestrians involved in the collision
PEDCYLCOUNT	The number of bicycles involved in the collision
VEHCOUNT	The number of vehicles involved in the collision
INCDATE	The date of the incident
INCDTTM	The date and time of the incident
JUNCTIONTYPE	Category of junction at which collision took place
SDOT_COLCODE	A code given to the collision by SDOT
SDOT_COLDESC	A description of the collision corresponding to the collision code
INATTENTIONIND	Whether or not a driver involved was under the influence of drugs or alcohol (Y/N)
UNDERINFL	A detailed description of the severity of the collision
WEATHER	A description of the weather conditions during the time of the collision
ROADCOND	The condition of the road during the collision
LIGHTCOND	The light conditions during the collision
PEDROWNOTGRNT	Whether or not the pedestrian right of way was not granted (Y/N)
SDOTCOLNUM	A number given to the collision by SDOT
SPEEDING	Whether or not speeding was a factor in the collision (Y/N)
ST_COLCODE	A code provided by the state that describes the collision
ST_COLDESC	A description that corresponds to the stateŠs coding designation
SEGLANEKEY	A key for the lane segment in which the collision occurred
CROSSWALKKEY	A key for the crosswalk at which the collision occurred
HITPARKEDCAR	Whether or not the collision involved hitting a parked car (Y/N)

Table 1: Dataset features. The feature SEVERITYCODE colored in red is the target feature.

severity code (i.e., the accident severity) by means of supervised machine learning algorithms. The other columns contain details about the accident such as:

- accident/report identifier (OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS);
- accident location (X,Y, INTKEY, LOCATION, SEGLANEKEY, CROSSWALKKEY);
- accident date and time (INCDATE, INCDTTM);
- type of collision (ADDRTYPE, COLLISIONTYPE, JUNCTIONTYPE, SDOT_COLCODE, SDOT_COLDESC, ST_COLCODE, ST_COLDESC);
- people and vehicles involved in the accident (PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, PEDROWNOTGRNT, HITPARKEDCAR);
- road condition (ROADCOND);
- light condition (LIGHTCOND);
- car speeding (SPEEDING);
- influence of drug/alcohol (UNDERINFL);
- lack of attention (INATTENTIONIND);
- further detailed description of the accident severity (EXCEPTRSNCODE, EXCEPTRSNDESC, SEVERITYCODE.1, SEVERITYDESC);

Not all the features are significant and negligible attributes for the aim of this investigation. Therefore the data has to be cleaned and wrangled before setting up a machine learning model.

2.2 Data cleaning

There are several problems with the datasets. First, the dataset contains different all the type of elements, integer float and object variables. Since the default type of index and columns is not a list, I get the index and columns as lists.

Second, not all the features are essential to the present investigation. The columns that refers to accident/report identifier (OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS), accident location (X,Y, INTKEY, LOCATION, SEGLANEKEY, CROSSWALKKEY), accident date and time (INCDATE, INCDTTM) are considered negligible for the aim of this investigation and thus dropped. Similarly, the columns that provides further detailed description of the accident severity

are considered redundant with the SEVERITY CODE which already sums up the accident severity. Therefore these features (namely, EXCEPTRSNCODE, EXCEPTRSNDESC, SEVERITYCODE.1, SEVERITYDESC) are dropped as well.

Third, among the features there are missing data that reaches also the 95% of the total number of observations. The missing data has to be handled properly. In features collecting categorical variables such as ADDRTYPE, COLLISIONTYPE, JUNCTIONTYPE, WEATHER, ROADCOND, LIGHTCOND the missing values are replaced by the most common value in the corresponding feature. Moreover, since the values 'Unknown' and 'Other' coexist in the features WEATHER, ROADCOND and LIGHTCOND, it is worth to replace the values labeled as 'Unknown' by the most common value in the corresponding feature. The highest percentage of missing values regards the features INATTENTIONIND, PEDROWNOTGRNT and SPEEDING which all contain binary variables as 'Y' and 'N'. Unfortunately, in the data collection only the values labeled as 'Y' were actually reported without filling in the database in case of value 'N'. For this reason the missing values in features INATTENTIONIND, PEDROWNOTGRNT and SPEEDING are replaced by value 'N'.

Next, there is some redundancy within some features. In the feature lighting condition (LIGHTCOND) the variables are grouped considering three light intensity: *'Daylight'* for high light intensity; *'Day/Night'* for medium light intensity grouping 'Dusk', 'Dawn' and 'Dark - Street Lights On', and *'Dark'* for low light intensity grouping 'Dark - No Street Lights', 'Dark - Street Lights Off' and 'Dark - Unknown Lighting'. The values referred as 'Other' and 'Unknown' are considered as missing values and are replaced by the most common value. Similarly, in the feature weather condition (WEATHER) the variables are grouped considering four weather forecast: *'Clear'* condition, *'Cloudy'* condition grouping 'Overcast' and 'Partly Cloudy', *'Windy'* condition grouping 'Fog/Smog/Smoke', 'Blowing Sand/Dirt' and 'Severe Crosswind', and *'Rain'* grouping 'Raining', 'Snowing' and 'Sleet/Hail/Freezing Rain'. Furthermore, in the feature weather condition (ROADCOND) the variables are grouped considering four slipperiness intensity: *'Dry'* for dry condition, *'Mixed'* for mixed condition grouping 'Sand/Mud/Dirt' and 'Snow/Slush', and *'Wet'* for wet condition grouping 'Wet', 'Icy', 'Oil' and 'Standing Water'. The values referred as 'Other' and 'Unknown' are considered as missing values and are replaced by the most common value. Moreover, in the feature regarding the influence of drug/alcohol as factor of accident severity (UNDERINFL), there are four variables: 'Y', '0', 'N' and '1'. Probably some mistakes were made in the data collection so that the actual influence of drug/alcohol was labeled both as 'Y' and '1' and the absence of use of drug/alcohol was labeled both as 'N' and '0'. These variables are thus grouped together so that the possible values of the feature UNDERINFL are just 'Y' or 'N'.

Droppedfeature	***Reason***
OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS, X,Y, INTKEY, LOCATION, SEGLANEKEY, CROSSWALKKEY, INCDATE, INCDTTM	Unnecessary features to assess the accident severity
EXCEPTRSNCODE, EXCEPTRSNDESC, SEVERITYCODE.1, SEVERITYDESC	Redundant features to assess the accident severity
COLLISIONTYPE, JUNCTIONTYPE, SDOT_COLCODE, SDOT_COLDESC, ST_COLCODE, ST_COLDESC	Redundant features with ADDRTYPE
PEDCOUNT, PEDCYLCOUNT, PEDROWNOTGRNT, HITPARKEDCAR	Redundant features with PERSONCOUNT and VEHCOUNT

Table 2: Dropped features recap.

Lastly, the features that have categorical variables are turned into numerical variables. For example the lighting intensity in the feature LIGHTCOND is converted as: 0 for low lighting intensity, 1 for medium light intensity and 2 for high lighting intensity. Similarly, the conversion is performed in features ADDRTYPE, ROADCON and WEATHER. Furthermore, the influence of drug/alcohol is converted as: 0 if drivers were not under influence of drug/alcohol and 1 if drivers were under influence of under influence of drug/alcohol. Similarly, the conversion is performed in features SPEEDING and INATTENTIONIND.

2.3 Feature selection

There is some redundancy among some feature. The type of collision is described by too many features ADDRTYPE, COLLISIONTYPE, JUNCTIONTYPE, SDOT_COLCODE, SDOT_COLDESC, ST_COLCODE and ST_COLDESC all correlated. The features ADDRTYPE synthetically describes the type of collision while all the other mentioned are dropped. Moreover, people and vehicles involved in the accident are described by too many features PERSONCOUNT,

PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, PEDROWNOTGRNT and HITPARKEDCAR all correlated. For sake of simplicity the features PERSONCOUNT and VEHCOUNT are kept in the database since they synthetically sum up number of people and vehicles involved in the accident while the other shown are dropped.

After cleaning the data and discarding the redundancies, the features are subdivided in two groups:

- independent variables (target): accident severity SEVERITYCODE;
- dependent variables (features): general type of collision ADDRTYPE, road condition ROADCOND, lighting condition LIGHTCOND, vehicle speeding SPEEDING, influence of drug/alcohol UNDERINFL, driver inattention, INATTENTIONIND, number of people involved in the accident PERSONCOUNT and number of vehicle involved in the accident VEHCOUNT;

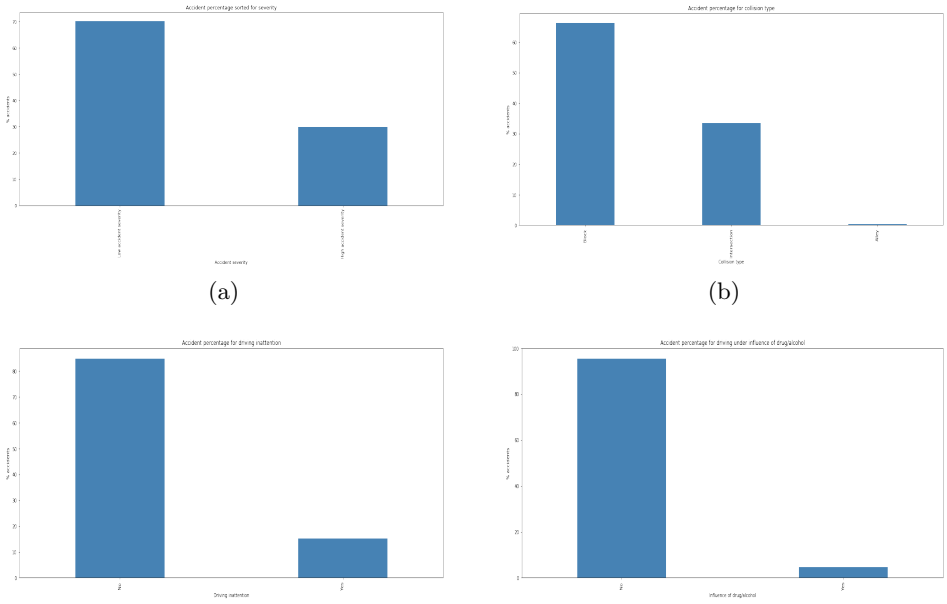
3 Data exploration

Before building the model, it is worth to have a look to the data. First, each variable is plotted in vertical bar to show the accident percentage (Figure 2 and **??**). The highest accident percentage has a low severity, namely with damage but no injury. Despite common perceptions, the accidents happen mostly during clear and sunny day, in dry road condition without exceeding the speed limit and without the influence of drug/alcohol. The majority of collisions happens without exceeding the speed limit. Sememingly most the accident are due to driving inattention occurring mostly at road blocks.

Further analysis can show how accident severity is related to group of several variables at time. Specifically, the relationship between the accident severity and other variables is plotted in heat maps which show the mean value of the accident severity[1] proportional to color with respect to the other variables in the vertical and horizontal axis. First, we would like to see the relationship between the mean accident severity, collision type and inattention at driving (Figure 4). The most severe accident takes place at intersections for driver inattention.

Second, we would like to see the relationship between the mean accident severity, lighting condition and road condition (Figure 5). The most severe accidents occur when the road is wet for all lighting conditions. Since the mean value of accident severity is high also when the road is dry for different lighting conditions, these two variables are seemingly not very interesting to find the most risky situations.

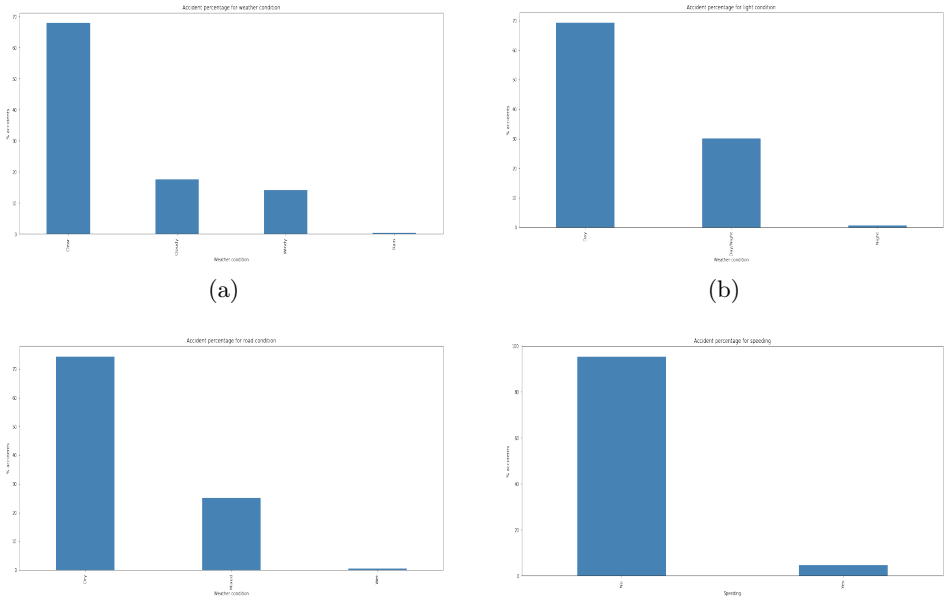
[1]The severity code which is a categorical variable which is equa to 1 for low severity accident or 2 for high severity accident. Computing a mean of a categorical variable just helps to see a behavior trend.



(a) (b)

(c) (d)

Figure 2: Accident percentage for a) severity code, b) collision type, c) inattention at driving and d) use of alcohol/drug.



(a) (b)

(c) (d)

Figure 3: Accident percentage for a) weather conditions , b) lighting conditions, c) road conditions and d) car speeding.

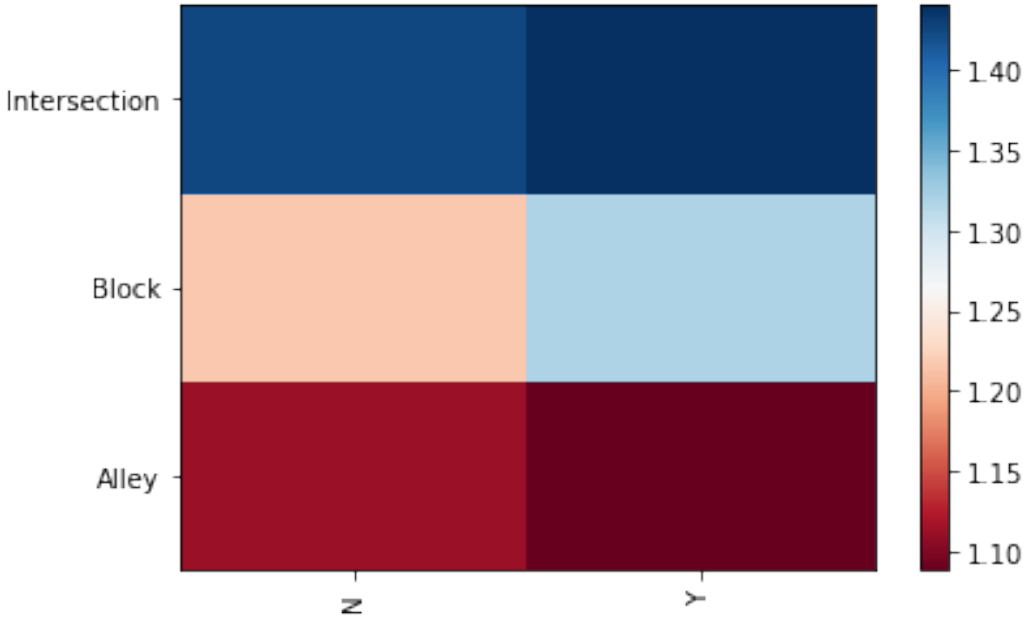


Figure 4: Relationship between mean accident severity changes, collision type (ADDRTYPE) and inattention at driving (INATTENTIONIND)

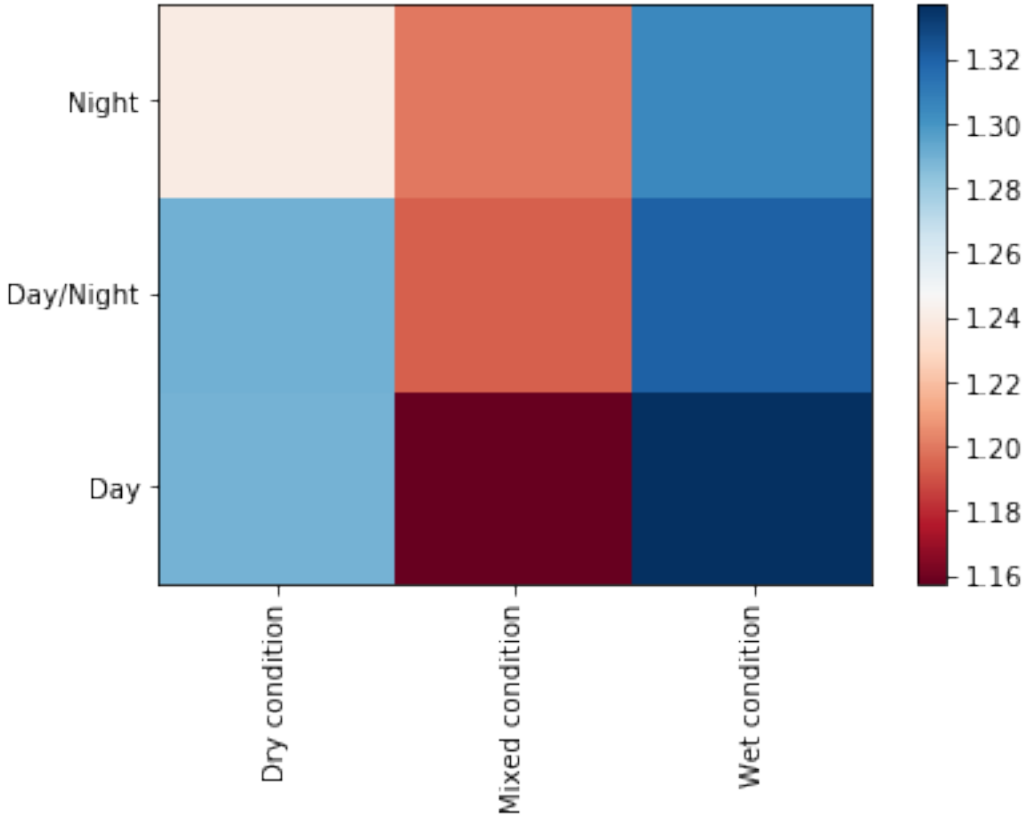


Figure 5: Relationship between mean accident severity changes, lighting condition (LIGHTCOND) and road condition (ROADCOND)

Third, we would like to see the relationship between the mean accident severity, weather condition and other variables (Figure 6). The rain and wind seem to increase the accident severity at daylight, in condition of high or medium visibility.

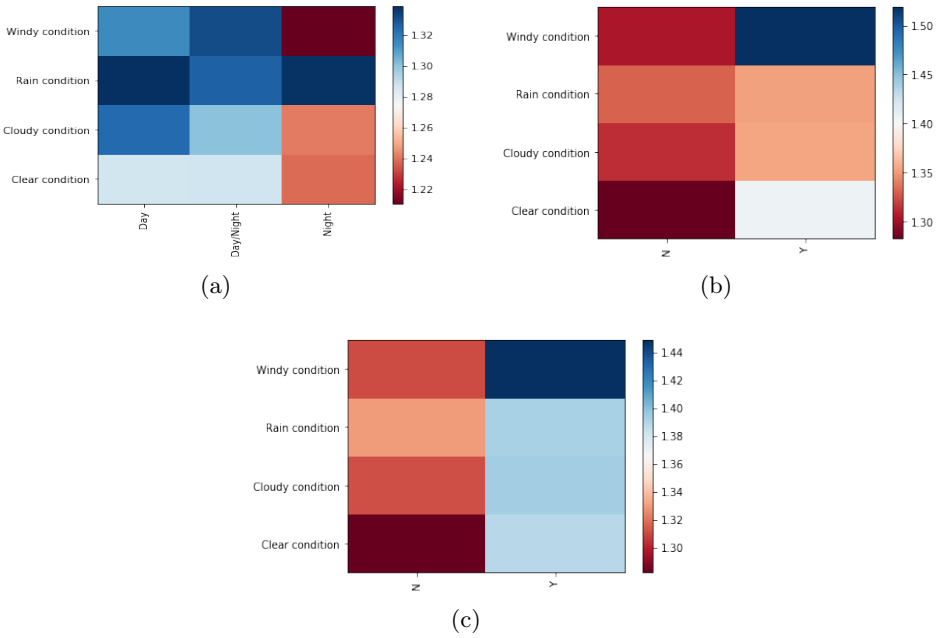


Figure 6: Relationship between mean accident severity changes, weather condition (WEATHER) and a) lighting condition (LIGHTCOND) , b) car speeding (SPEEDING), c) influence of drug/alcohol (UNDERINFL).

In the last visuals it is not clear which feature impacts the most on the severity code. Determining the factors affecting the most the accident severity is performed by correlation analysis. The Pearson Correlation measures the linear dependence between two variables X and Y. The resulting coefficient is a value between -1 and 1 inclusive, where:

- 1: Total positive linear correlation;
- 0: No linear correlation, the two variables most likely do not affect each other
- -1: Total negative linear correlation.

The P-value is the probability value that the correlation between these two variables is statistically significant. Normally, the tagret level of significance is 0.05, which means the correlation between the variables is significant with a confidence level of 95%. By convention, P-value is subdivided in four subsets:

- P-value < 0.001: there is strong evidence that the correlation is significant;

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

Table 3: The values of Pearson coefficient and P-value of accident severity (SEVERITYCODE) and other variables.

- P-value < 0.05: there is moderate evidence that the correlation is significant;
- P-valus < 0.1: there is weak evidence that the correlation is significant;
- P-value > 0.1: there is no evidence that the correlation is significant.

Computing the Pearson coefficient and the P-value for the numerical features, we see that both are correlated to accident severity with a positive relationship (Table 3). The more severe is the accident, the higher is the number of people and vehicles involved in the accident.

The study of the correlation between categorical variables is done by the analysis of variance. The Analysis of Variance (ANOVA) is a statistical method used to test whether there are significant differences between the means of two or more groups. ANOVA returns two parameters:

- F-test score, ANOVA assumes the means of all groups are the same, calculates how much the actual means deviate from the assumption, and reports it as the F-test score. A larger score means there is a larger difference between the means;
- P-value, level of significance of F-test score;

In Table 4 are shown how different categorical variables impact on the accident severity. Each features has a significant correlation with the accident severity (low level of P-value). Furthermore, the features have a strong impact on the accident severity (high value of F-test score) with the exception of lighting condition (low value of 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).

4 Predictive modeling

4.1 Pre-processing

Once defined, the features are subjected to standardization since variables are measured with different units and at different scales. Features and target are subdi-

	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	$1.908 \cdot 10^{-85}$

Table 4: ANOVA results: the values of F-score and P-value of accident severity (SEVERITYCODE) and other variables.

vided in two subsets: one for training the data (80%) and the other for testing the data (20%). This test/train split enables a proper use of the available data.

4.2 Classification algorithms

The labeled data accident severity is used to train and test a model by supervised machine learning. The machine learning algorithm considered are:

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

Not all the classification algorithms scale easily with the dataset size. Specifically, KNN and SVM do not scale easily. *The time complexity for training K-Nearest Neighbors Training* $O(1)$ *for training which means it is constant and O(n) for testing* (further details about the complexity of KNN algorithm at link). *The fit time complexity is more than quadratic with the number of samples which makes it hard to scale to dataset with more than a couple of 10000 samples. For the linear case, the algorithm used in LinearSVC by the liblinear implementation is much more efficient than its libsvm-based SVC counterpart and can scale almost linearly to millions of samples and/or features* (further details about the complexity of SVC algorithm at link).

4.3 Model evaluation

Model performance is assessed by Jaccard index F1-score and, if applicable, by log-loss index (Table 5).

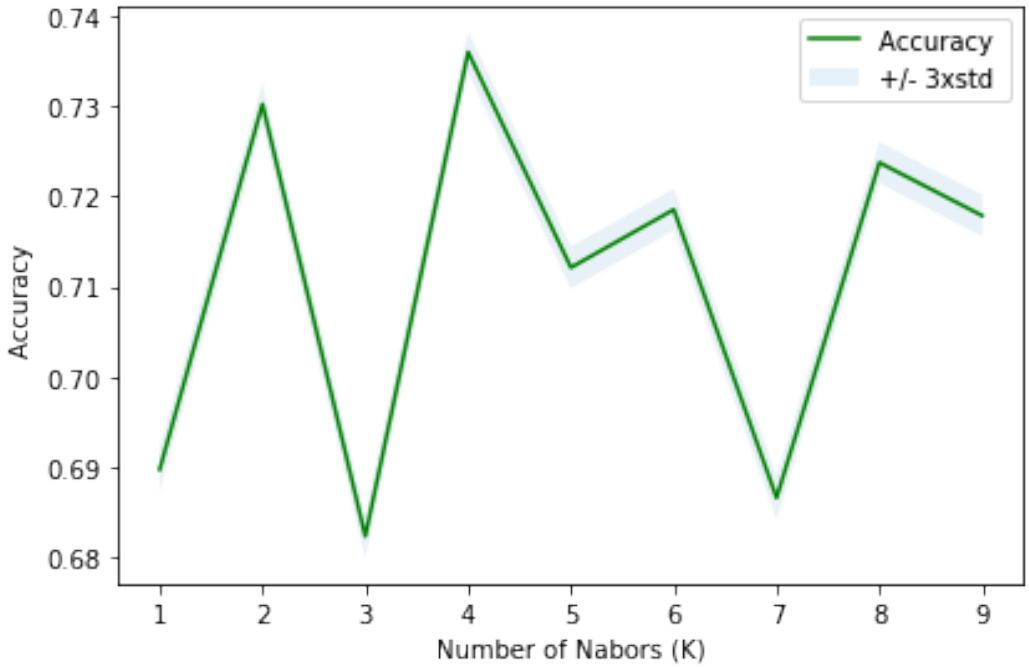


Figure 7: The K-Nearest Neighbors accuracy for different k values.

	$Jaccard$	$F1-score$	$Log-loss$
KNN	0.7	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

Table 5: Model evaluation

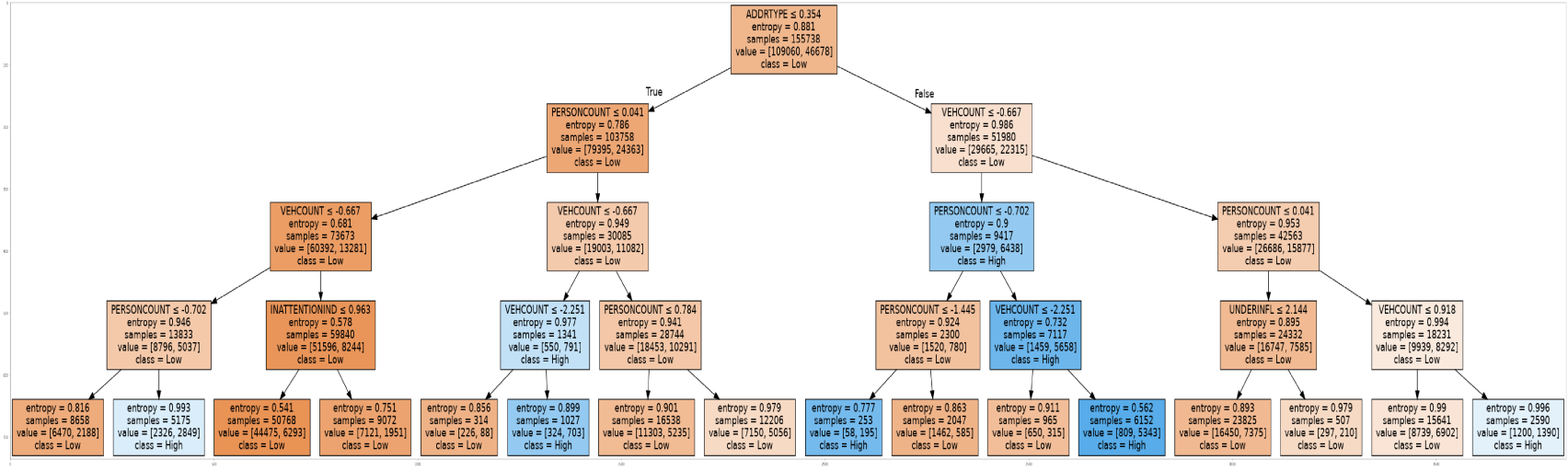


Figure 8: The decision tree.

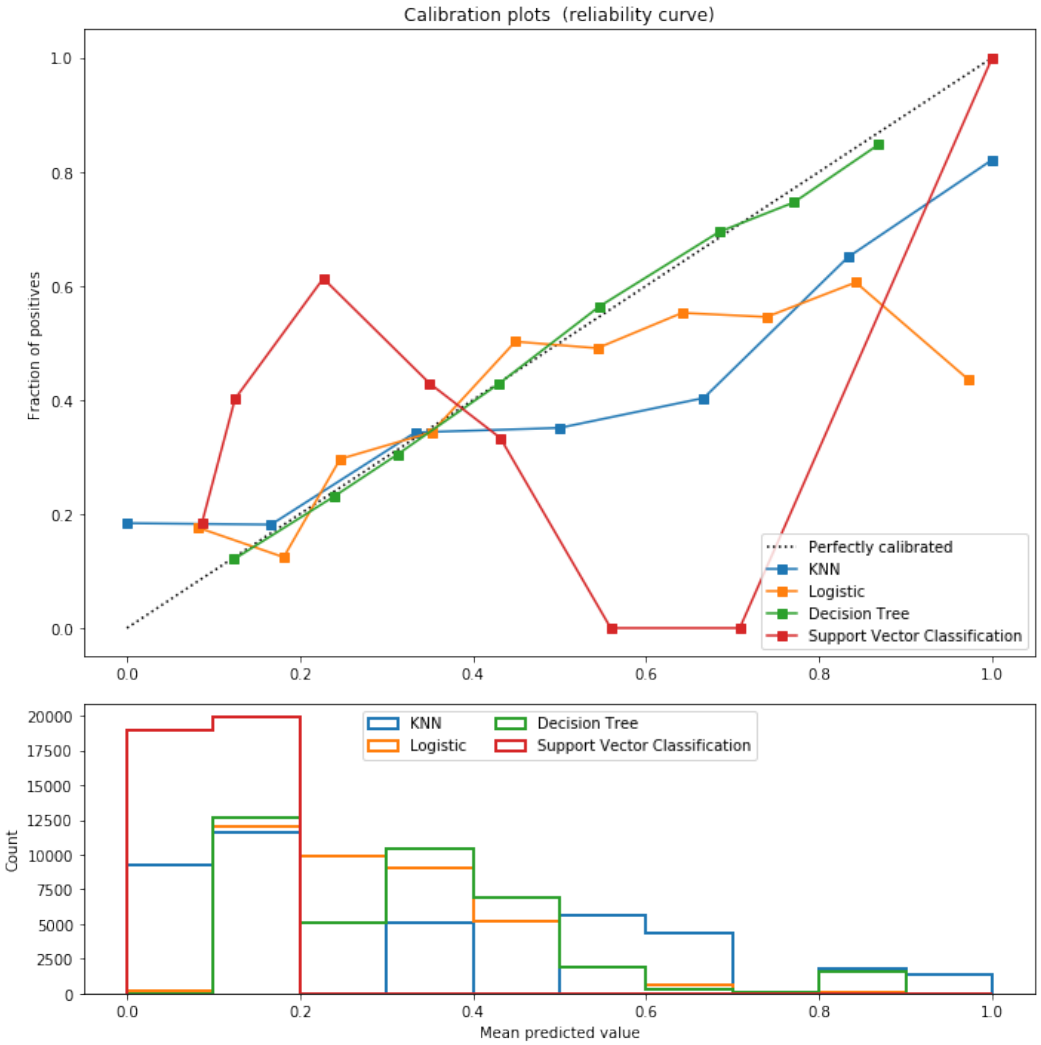


Figure 9: Calibration curve

5 Conclusions

This study aims at predicting the severity of an accident. The data source refers to accidents record occurred in Seattle (US) from 2004 onwards. . After cleaning the data, the target consists in the severity code, a number representing the severity of an accident ranging for the present dataset from 1, low severity accident, to 2, high severity accident. The features refer to type of collision, weather conditions, lighting conditions, road conditions, car speeding, influence of alcohol/drug, driving inattention, number of people and vehicle involved in the accident.

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 fro 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.