Data Science: A Road to Safer Roads

By

Dr Debdarsan Niyogi

10th September, 2020

Contents

Introduction

According to the statistics by WHO (7th Feb, 2020):

- Every year the lives of approximately 1.35 million people are cut short as a result of a road traffic crash. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury.
- Road traffic injuries cause considerable economic losses to individuals, their families, and to nations as a whole. These losses arise from the cost of treatment as well as lost productivity for those killed or disabled by their injuries, and for family members who need to take time off work or school to care for the injured. Road traffic crashes cost most countries 3% of their gross domestic product.
- Road traffic injuries are the leading cause of death for children and young adults aged 5-29 years.

This, therefore, needs serious attention, as it concerns human lives which is *irreplaceable*. It is possible, thanks to machine learning, to predict severity of car accidents as a result of the complex interplay of multitudes of factors like weather, road condition, light condition, speeding etc. and also to identify which factors are more important. The information thus gathered can be used to take preventive measurements.

Objective

In 2017, WHO released *Save LIVES (a road safety technical package* focuses on **S**peed management, **L**eadership, **I**nfrastructure design and improvement, **V**ehicle safety standards, **E**nforcement of traffic laws and post-crash **S**urvival*) which* synthesizes evidence-based measures that can significantly reduce road traffic fatalities and injuries.

While the pursuit of saving lives is obvious, the 2030 Agenda for Sustainable Development has set an ambitious and *quantifiable* goal of halving the global number of deaths and injuries from road traffic crashes by 2020.

According to the National Safety Council, traffic collisions cause more than 40,000 deaths and injure thousands of people every year across the United States. *These are not traffic accidents, but entirely preventable tragedies.*

In order to reduce accidents, we need to predict it based on the external parameters. Since the accident occurs due to very many factors (unsafe road infrastructure, light condition, vulnerable road users, speeding, driving under the influence of alcohol and other psychoactive substances, distracted driving, weather) prediction of accidental severity is a challenge. Machine Learning is ideally suited here as this is a scientific approach for modelling and predicting the parameter of interest demanding only a low budget.
The current project attempts to apply a machine learning technique to predict the severity of the accident given the parameters as stated before using car collision data for the city of Seattle, USA. Additionally, the accident data is extremely skewed (class imbalance), which makes the models heavily biased towards mild injuries. A practical methodology is used here to tackle this situation.

Interests

The practical utilities of the prediction, besides saving lives:

- Safe route planning
- Emergency vehicle allocation
- Roadway design
- Reduce property damage
- Where to place additional signage (e.g. to warn for curves)

Study of accidents in one US city can definitely help any other city having similar characteristics. As car accidents are correlated to socioeconomic condition (according to WHO), the model developed cannot be used across countries but the same methodology can, of course, be utilised. The stakeholders of the present problem are federal, state and local government agencies, non-governmental organizations, regional authorities, and possible individuals (if this model is deployed as an app for personal use).

Data

The car collision data is obtained from Seattle Govt's website (Timeframe: 2004 to Present). The meaning of the data attributes (features) can be found here.

Upon loading the data, we see that there are 40 variables and 2,21,006 number of observations with zero duplicates but having a missing cell percent 15.8%.

Table 1: Dataset Statistics

Number of variables	40
Number of observations	221006
Missing cells	1393590
Missing cells (%)	15.80%
Duplicate rows	0
Duplicate rows (%)	0.00%
Total size in memory	338.6 MiB
Average record size in memory	1.6 KiB

Table 2: Variable types

CAT	23
NUM	16
BOOL	1

We see that the Severity level of the car accidents to be predicted (*SEVERITYCODE*) are listed against 39 independent variables (features), like weather, light condition, road condition, date and time of the accident, collision codes, collision type, collision address type, types of injuries, junction type where the accident took place etc. The task is to predict the accident severity given all other factors or features.

Some important features are summarized below to get a feeling of what we are dealing with:

Table 3: Important Features

Attribute	**Description**
ADDRTYPE	Collision address type: Alley, Block, Intersection
SEVERITYCODE	A code that corresponds to the severity of the collision: 3—fatality, 2b—serious injury, 2—injury, 1—prop damage, 0—unknown
PERSONCOUNT	The total number of people involved in the collision
PEDCOUNT	The number of pedestrians involved in the collision. This is entered by the state.
PEDCYLCOUNT	The number of bicycles involved in the collision. This is entered by the state.
VEHCOUNT	The number of vehicles involved in the collision. This is entered by the state.
INJURIES	The number of total injuries in the collision. This is entered by the state.
SERIOUSINJURIES	The number of serious injuries in the collision. This is entered by the state.
FATALITIES	The number of fatalities in the collision. This is entered by the state.
INCDTTM	The date and time of the incident.
JUNCTIONTYPE	Category of junction at which collision took place
INATTENTIONIND	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.
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.
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. For more information about these codes, please see the State Collision Code Dictionary.
ST_COLDESC	A description that corresponds to the state's coding designation.
HITPARKEDCAR	Whether or not the collision involved hitting a parked car. (Y/N)

Initial Data Processing

Now we are ready to proceed towards data preparation methodically.

Date-Time Variable

To explore the date-time feature, let us extract the Month, Weekday, Hour information from *INCDTTIME* and add those as separate columns.

Dropping Useless Variables

A few variables are discarded because of the following factors:

- 1. There is no definition/description of the attribute (unknown data)

- 2. Unique keys/numbers
- 3. Redundant data, for example, both incident date and incident date-time are present. Other examples are having both code and corresponding description columns, where the description column containing the definitions of code values are superfluous.

A table is presented here to describe which features are eliminated and for what reason:

Table 4: Features dropped

Attribute	**Description**	**Reason for Dropping**
INCDATE	The date of the incident.	INCDTTM has both date and time info
INCDTTM	The date and time of the incident.	Extracted Year, Month, Day info
OBJECTID	ESRI unique identifier	Unique identifier is not a predictor
INCKEY	A unique key for the incident	Unique identifier is not a predictor
COLDETKEY	Secondary key for the incident	Unique identifier is not a predictor
INTKEY	Key that corresponds to the intersection associated with a collision	Unspecified key
SEGLANEKEY	A key for the lane segment in which the collision occurred.	Unspecified key
CROSSWALKKEY	A key for the crosswalk at which the collision occurred.	Unspecified key
EXCEPTRSNCODE	Not specified	Unknow key, where the only value is 'Not Enough Information (NEI)
REPORTNO	Unknow key	Unknow key
X	Unknown	Unknown key
Y	Unknown	Unknow key
STATUS	Unknow key	Unknow key
SDOTCOLNUM	A number given to the collision by SDOT.	Unknow Number or Code
EXCEPTRSNDESC	Not specified	Description is not needed
SEVERITYDESC	A detailed description of the severity of the collision	Description is not needed; we have the key
SDOT_COLDESC	A description of the collision corresponding to the collision code.	Description is not needed; we have the key
SDOT_COLCODE	Collision code assigned by SDOT	We already have ST_COLCODE, and this key is not explained
ST_COLDESC	A description that corresponds to the state's coding designation.	Description is not needed; we have the key
LOCATION	Description of the general location of the collision	Description is not needed (we do not have corresponding codes)

Exploratory Data Analysis (EDA)

Now our data is relatively cleaner and lighter. In this phase we concentrate on *Data visualization, Cleaning up missing data, Value imputation, Value re-grouping*.

Severity Codes

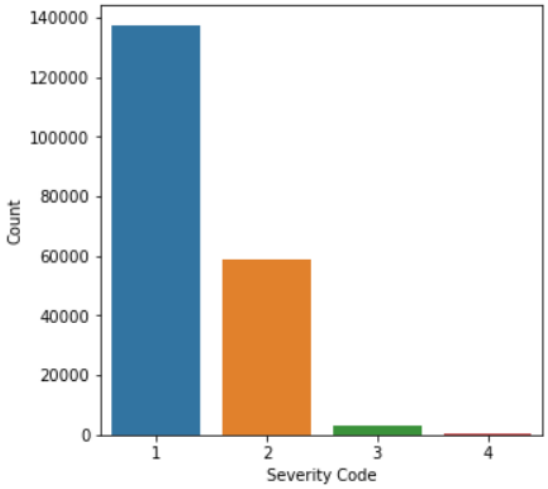
Severity Code (SEVERITYCODE) is the target/dependent variable. Let us scrutinize that first. Accident Severity is numbered from 0 to 3. We inspect the percentages of severity codes:

Table 5

Severity Code	% Count
1	62.14
2	26.53
0	9.77
2b	1.40
3	0.16

Since we are not going to predict an 'Unknown' severity (*SEVERITYCODE* = 0), these observations, along with the rows with missing values can safely be deleted. The categorical values can also be remapped to a scale of 1 to 4, where 3 is assigned to 4 and 2b to 3. The new mapping is as follows:

SEVERITYCODE	SEVERITYDESC
1	Property Damage Only Collision
2	Injury Collision
3	Serious Injury
4	Fatal Collision

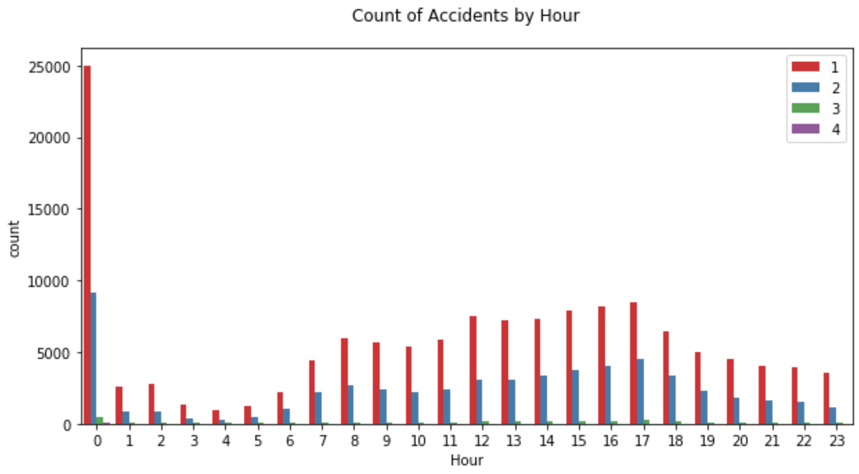


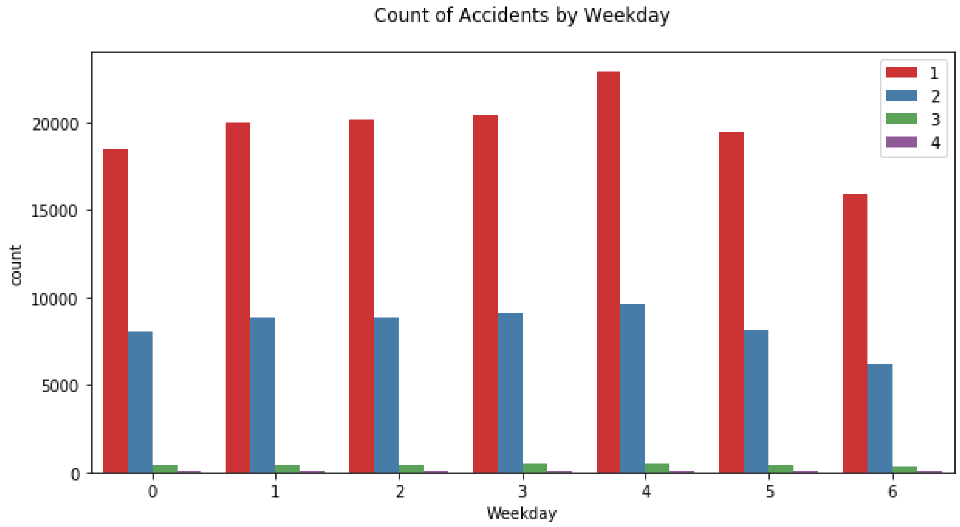
This is to be noted that the highest severity (4 = Fatal) has only 0.18% of the number of observations, in other words, the distribution is extremely skewed:

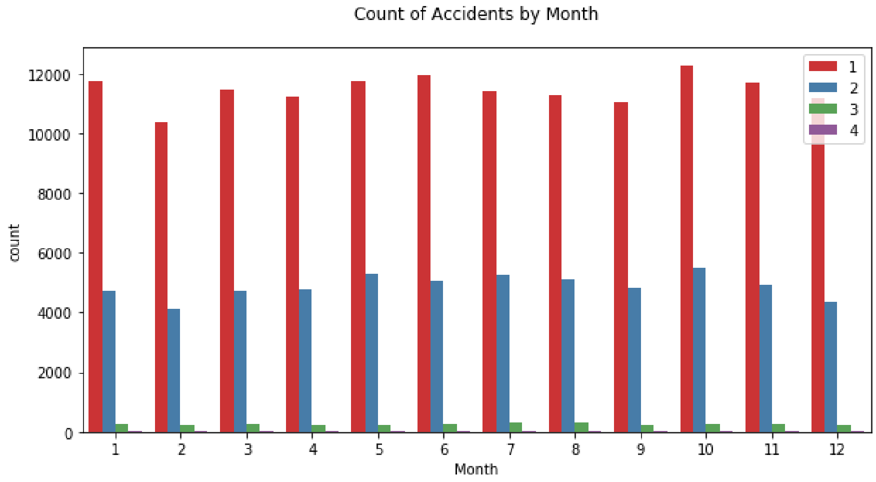
As expected, Severity count decreases rapidly with increase in severity. We see that the count of severity 3 and 4 are very less.

Now that the target variable is inspected, we analyze each predictor features, one by one, and take the required measures (tackle missing values, value imputation and re-grouping) as we go along.

Date-time features



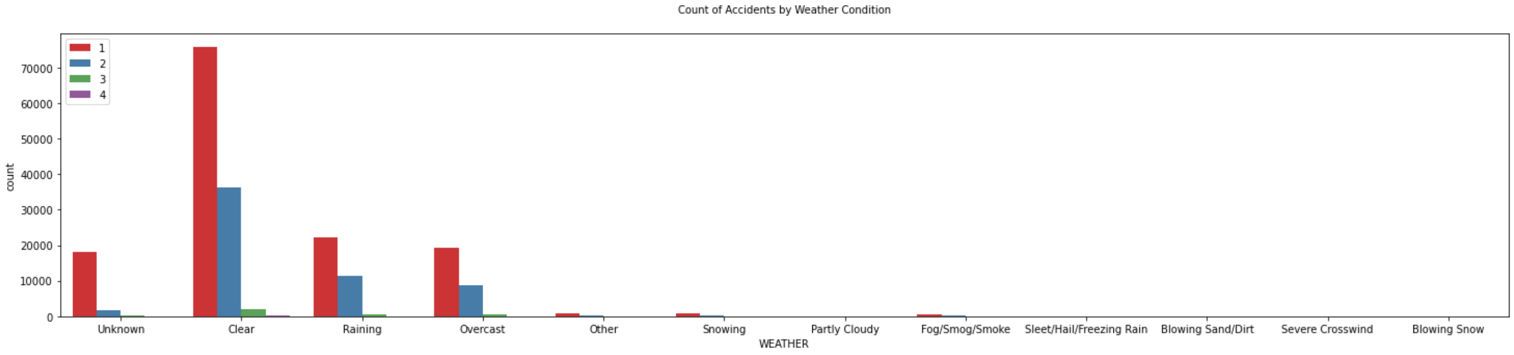




Weather features

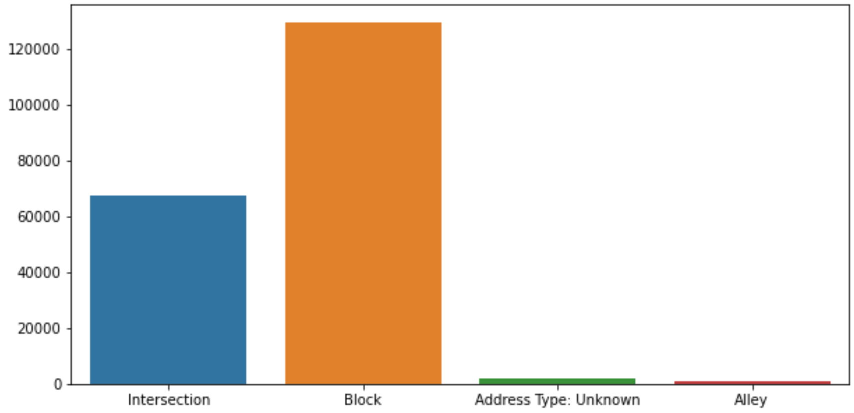
The blank (or NaN) values are substituted by an already existing category 'Unknown'. The percentage of occurrences are as follows:

Weather	% Count
Clear	57.32%
Raining	17.05%
Overcast	14.29%
Unknown	10.05%
Snowing	0.46%
Other	0.43%
Fog/Smog/Smoke	0.29%
Sleet/Hail/Freezing Rain	0.06%
Blowing Sand/Dirt	0.03%
Severe Crosswind	0.01%
Partly Cloudy	0.01%
Blowing Snow	0.00%

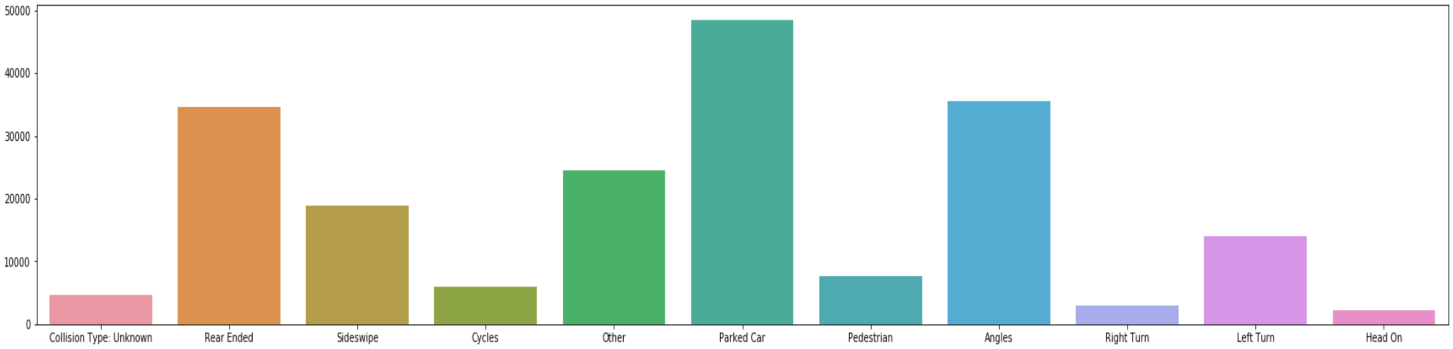


Collision Address Type

This denotes if the collision has taken place in a block, intersection or alley. After replacing the missing values with 'Unknown', we get the following diagram:

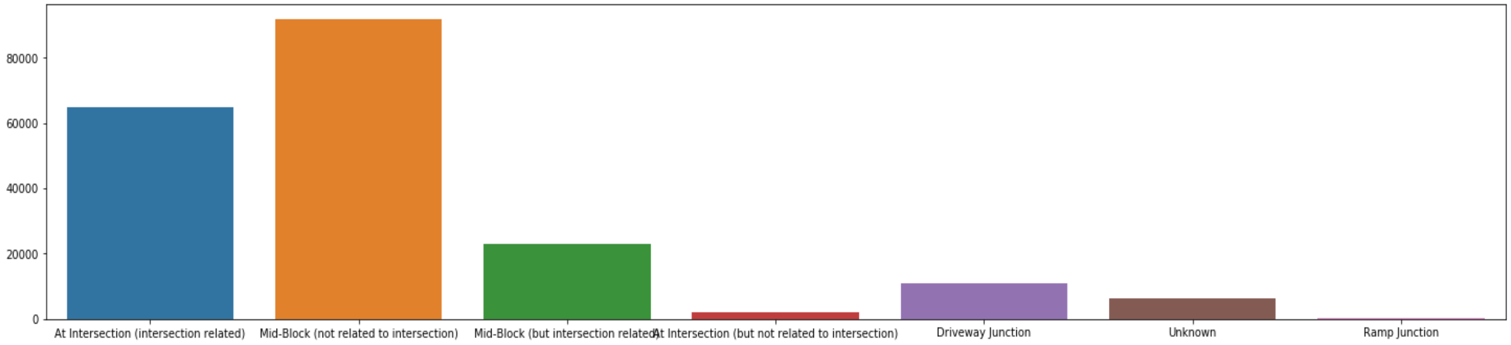


Collision Type



Here we have substituted NaN with 'Unknown'.

Junction Type

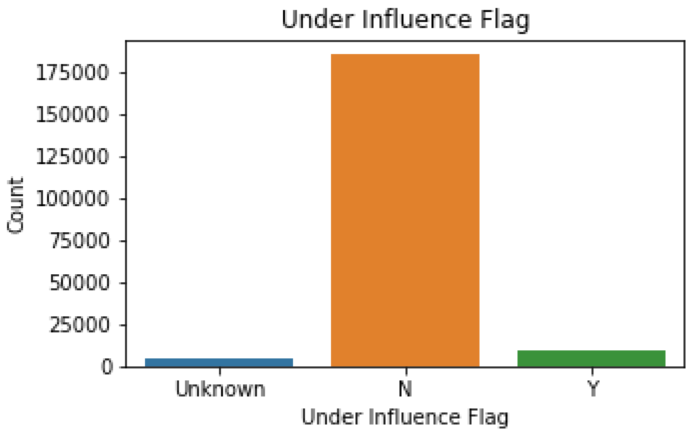


Junction Type	% Count
Mid-Block (not related to intersection)	46.07%
At Intersection (intersection related)	32.57%
Mid-Block (but intersection related)	11.56%
Driveway Junction	5.44%
NaN	3.20%
At Intersection (but not related to intersection)	1.07%
Ramp Junction	0.09%
Unknown	0.01%

Like before, we substitute the missing values with 'Unknown'.

Under Influence

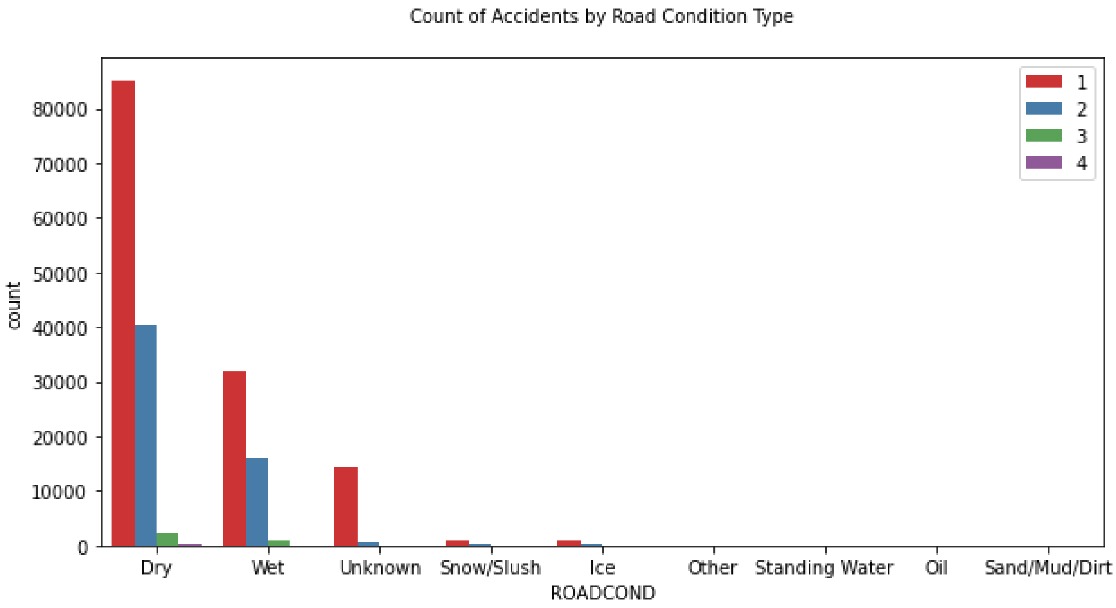
This flag indicates whether or not a driver involved was under the influence of drugs or alcohol.



We see that there are ambiguous data and even if we merge '0' label with 'N' and '1' with 'Y'.

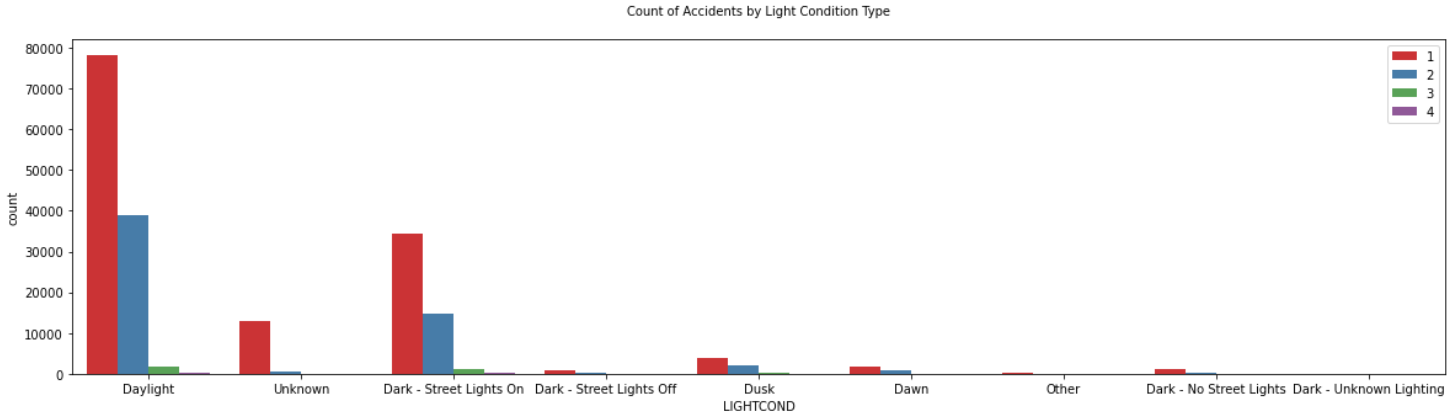
Road Condition

We substitute NaN by an already existing category 'Unknown' and plot the values to see that the accidents occurred mostly in dry road condition, which is counter-intuitive.



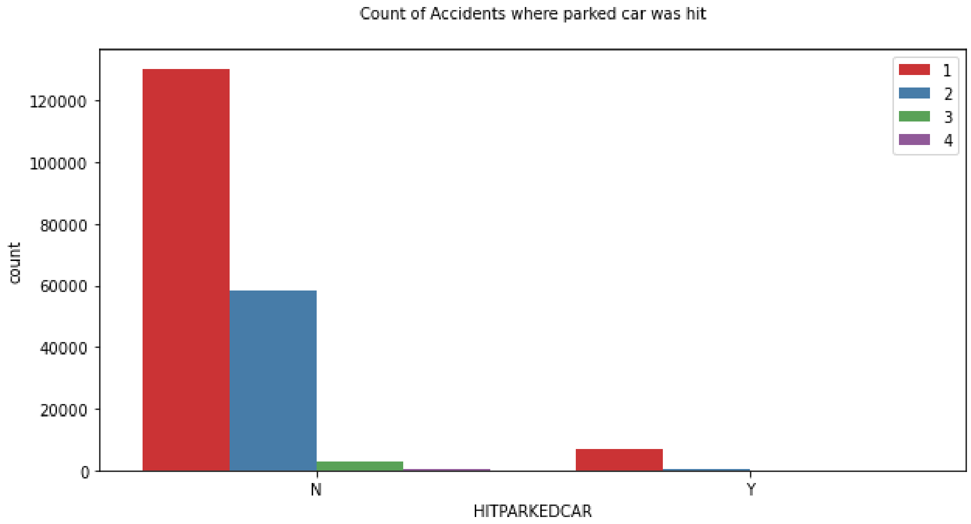
Light Condition

This is processed exactly as Road Condition, and we see that maximum accidents happened during daylight.



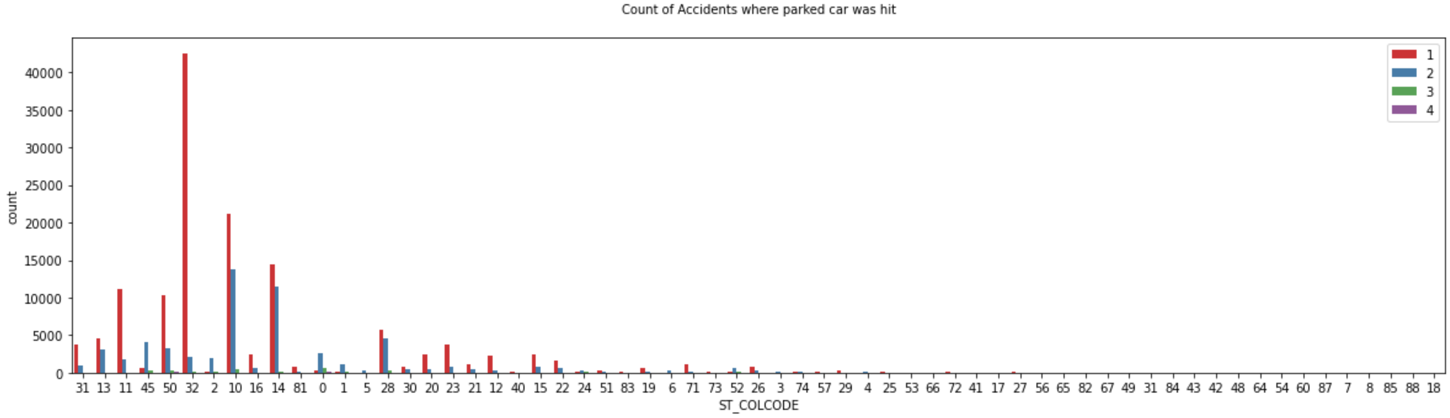
Hit Parked Car

This flag is set to 'Y' if the collision involved hitting a parked car.



Collision Codes

The missing values of ST_COLCODE are substituted by code 31 ('Not Stated'), and the resulting plot is below:



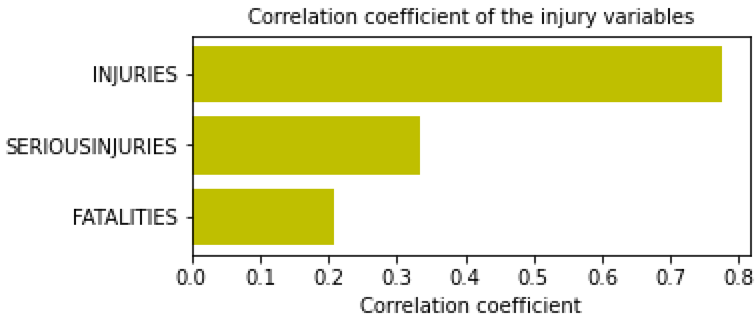
We see that the top codes are 32 (One Parked - One Moving), 10 (Entering At Angle) and 14 (From Same Direction - Both Going Straight – One Stopped - Rear End).

Injuries, Serious Injuries and Fatalities Features:

There are three variables: Injuries, Serious injuries and Fatalities. Let us see how the numbers are distributed amongst the four severity codes we have.

SEVERITYCODE	sum_INJURIES	sum_SERIOUSINJURIES	sum_FATALITIES
1	0	0	0
2	77939	0	0
3	4396	3248	0
4	288	103	372

The matrix is indicating a very strong correlation with severity.



Since, apart from Severity Code = 1 ("Property Damage Only Collision"), Severity code is assigned based on the injury level, and so the former is a direct reflection of the latter. If we use injury features as predictors, it is easily seen that those will overwhelm the other features, and the prediction will be based on the after-effects of a collision. Therefore, these three features are dropped.

The Rest of the Variables

Pedestrian ROW (Right Of Way) not granted, Speeding, Inattention indicator (PEDROWNOTGRNT, SPEEDING and INATTENTIONIND) Variables. These variables have missing values which are substituted by 'Unknown'.

The variables Person Count, Pedestrian Count, Pedestrian Cycle Count and Vehicle Count (PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT) do not have missing values as seen in the profile report above.

We have now come to the end of data analysis.

Mode Analysis

It will be interesting to see the mode (highest frequency) values of each feature with respect to the severity codes (where **S** indicates severity code).

Feature	Mode (S = 1)	Mode (S = 2)	Mode (S = 3)	Mode (S = 4)
COLLISIONTYPE	Parked Car	Rear Ended	Pedestrian	Pedestrian
PERSONCOUNT	2	2	2	2
PEDCOUNT	0	0	0	0
PEDCYLCOUNT	0	0	0	0

VEHCOUNT	2	2	1	1
INJURIES	0	1	1	1
SERIOUSINJURIES	0	0	1	1
FATALITIES	0	0	0	0
JUNCTIONTYPE	Mid-Block (Not Intersect)	At Intersection (Intersect)	At Intersection (Intersect)	At Intersection (Intersect)
INATTENTIONIND	Y	Y	Y	Y
UNDERINFL	N	N	N	N
WEATHER	Clear	Clear	Clear	Clear
ROADCOND	Dry	Dry	Dry	Dry
LIGHTCOND	Daylight	Daylight	Daylight	Daylight
PEDROWNOTGRNT	Y	Y	Y	Y
SPEEDING	Y	Y	Y	Y
ST_COLCODE	32	10	0	0
HITPARKEDCAR	N	N	N	N
Month	10	10	7	7
Weekday	4	4	4	4
Hour	0	0	0	0

Feature Selection

Let us select the features of interest, and one-hot encode them.

- ADDRTYPE
- COLLISIONTYPE
- JUNCTIONTYPE
- INATTENTIONIND
- UNDERINFL
- WEATHER
- ROADCOND
- LIGHTCOND
- PEDROWNOTGRNT
- SPEEDING
- ST_COLCODE
- HITPARKEDCAR

In the process of one-hot encoding of the category features described above, we have created columns with headings containing 'Unknown' for almost each of those. We can safely get rid of those columns as they convey hardly any meaning.

Classification Models for Multi-class, Skewed Distribution

Although we know *a priori*, that this extremely skewed and multi-class data may not be amenable even to the specialized classification models that deal with unbalanced data, we go ahead and have a taste of their performance, nevertheless. Here we have chosen

a) Bagging

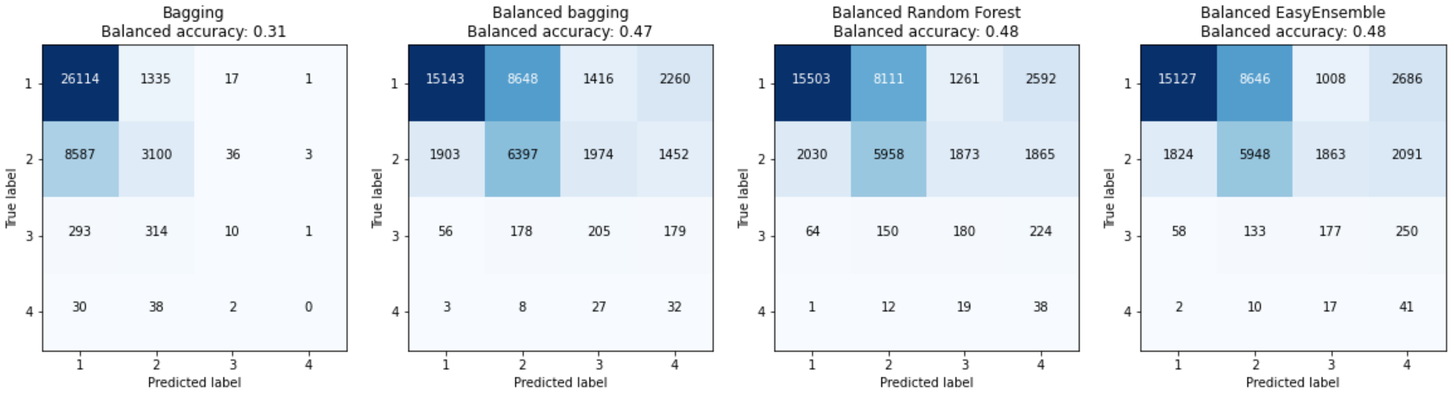
b) Balanced Bagging

c) Balanced Random Forest and

d) EasyEnsemble

classifiers that are capable to deal with unbalanced data inherently.

The confusion matrices generated for the above models:



Bagging
Balanced accuracy: 0.31

True label \ Predicted label	1	2	3	4
1	26114	1335	17	1
2	8587	3100	36	3
3	293	314	10	1
4	30	38	2	0

Balanced bagging
Balanced accuracy: 0.47

True label \ Predicted label	1	2	3	4
1	15143	8648	1416	2260
2	1903	6397	1974	1452
3	56	178	205	179
4	3	8	27	32

Balanced Random Forest
Balanced accuracy: 0.48

True label \ Predicted label	1	2	3	4
1	15503	8111	1261	2592
2	2030	5958	1873	1865
3	64	150	180	224
4	1	12	19	38

Balanced EasyEnsemble
Balanced accuracy: 0.48

True label \ Predicted label	1	2	3	4
1	15127	8646	1008	2686
2	1824	5948	1863	2091
3	58	133	177	250
4	2	10	17	41

Confusion matrices along with balanced accuracy for different models

We see that in all the four cases, balanced accuracy did not even cross 50%.

Multi-class to Two-class

Often, the skewed multi-class classification problem is converted to the two-class problem by taking the minority class versus the group of the rest of the classes. In our situation, the accidents with severity level 4 are fatal and others are non-fatal. Therefore, we can focus on level 4 accidents and regroup the levels of severity into level 4 versus other levels. In this process, a new column 'Severity 4' is created.

Balancing the Data

As seen above, Severity 4 is extremely rare, or in other words, the data is highly skewed. The main challenge of dealing with this type of data is that the machine learning algorithms train with almost 100% accuracy and fails to classify the minority class. This is intuitive since when the occurrence of the majority class is 99% per cent, even if the classifier is hard-coded to predict majority class always, the accuracy will still be 99%.

We appreciate that false negative is very costly here, that is actual severity code 4 is not predicted. The situation is just like the detection of fraudulent transactions or diagnosing diseases. There are many ways to deal with this situation by balancing the data synthetically by exploration method before training. We might

(1) under-sample the majority class

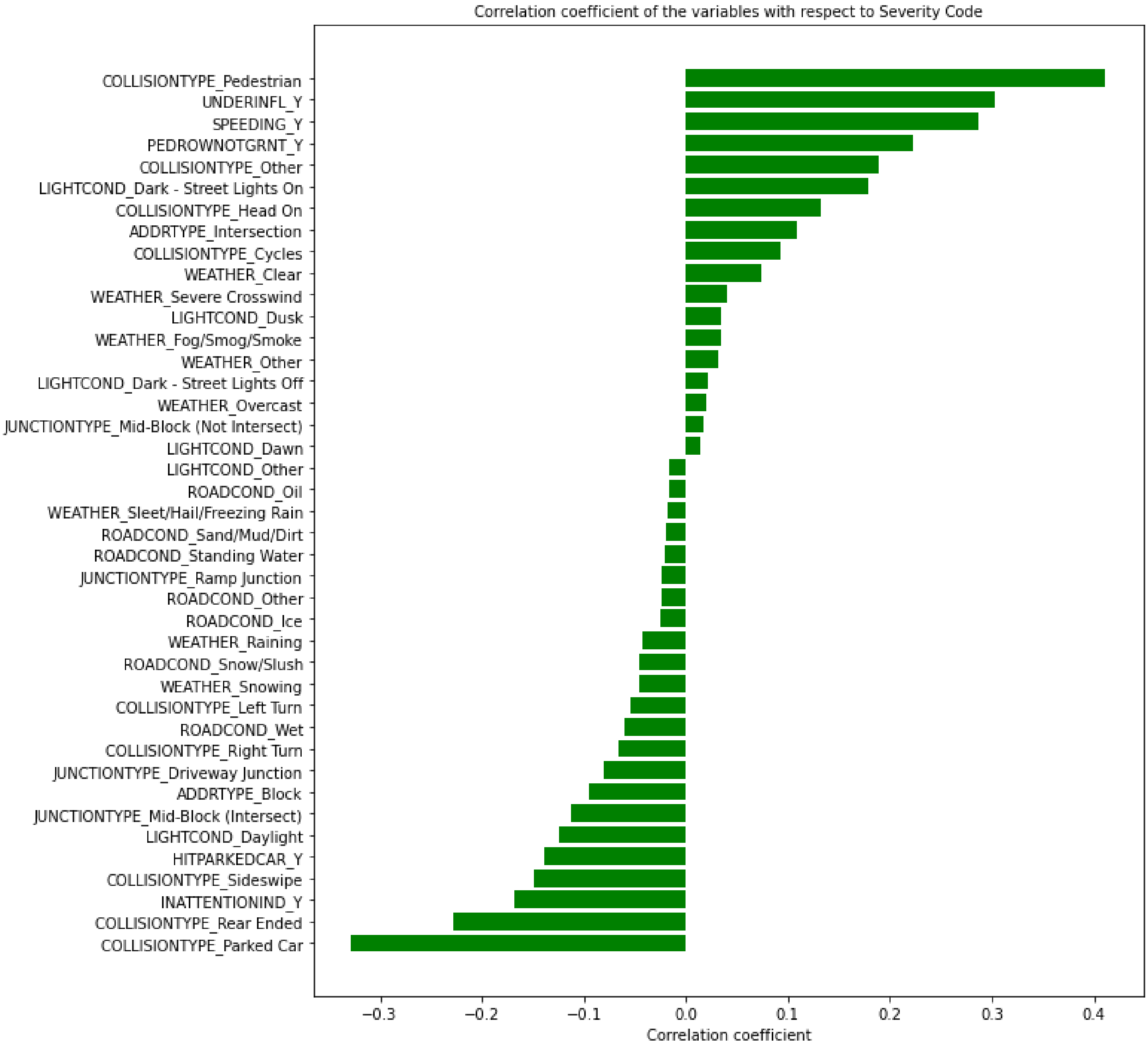
(2) over-sample the minority class or

(3) have a combination of (1) and (2), i.e. over- and under-sample simultaneously.

The combination of over- and under-sampling will be used since the data is large enough. level 4 will be randomly over-sampled to 10000 and other levels will be randomly under-sampled to 10000.

Correlation Coefficients

Let us now get an idea of how the variables are correlated (except ST_COLCODE, as it contains a very long list of codes).



We can see that the variables are not highly correlated.

Classification Models (applied to balanced data)

We are going to consider:

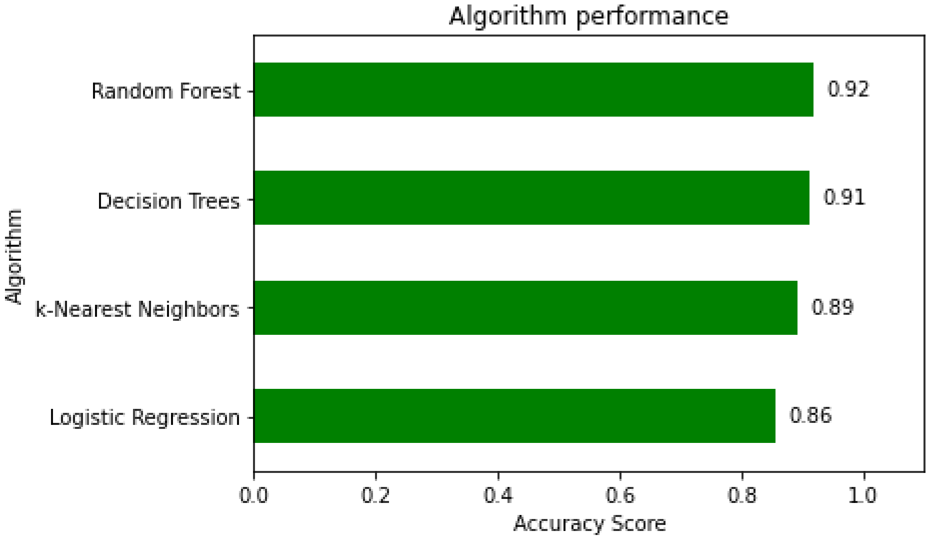
a) Logistic Regression,
b) k-Nearest Neighbors (kNN),
c) Decision Tree
d) Random Forest

models.

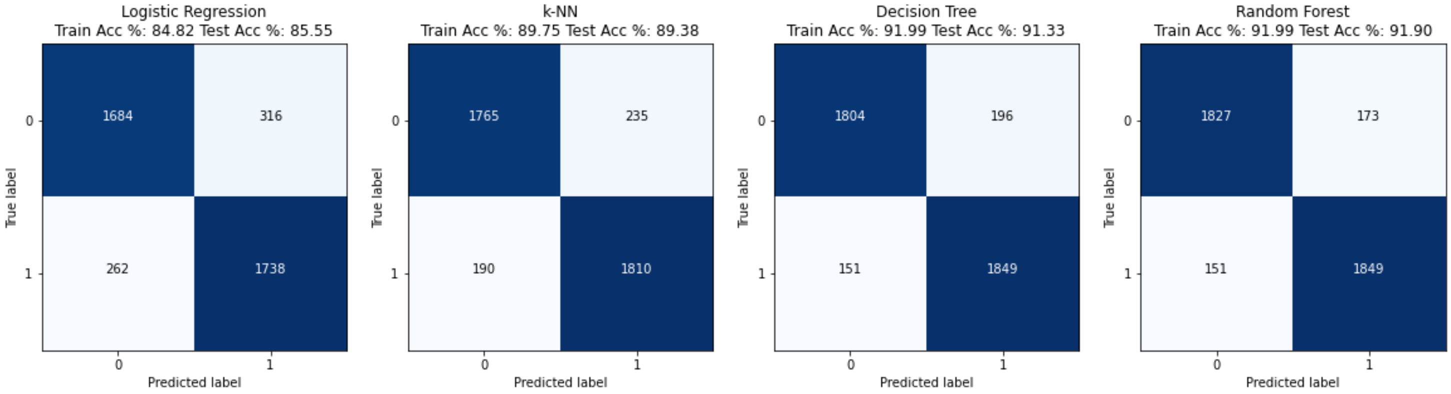
The above four model results are summarized below.

Accuracy

Accuracies achieved by different algorithms are shown here.

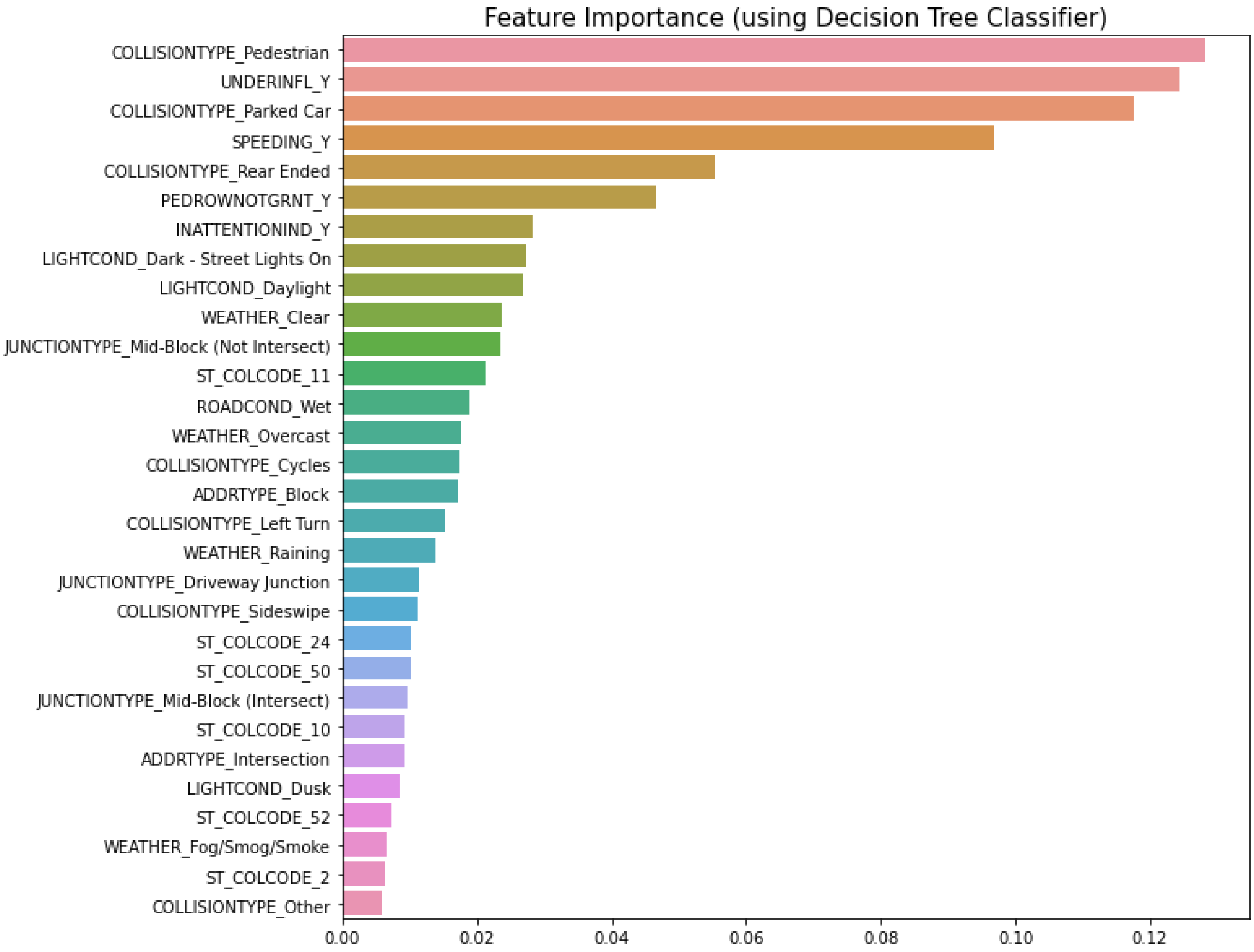


Confusion matrices

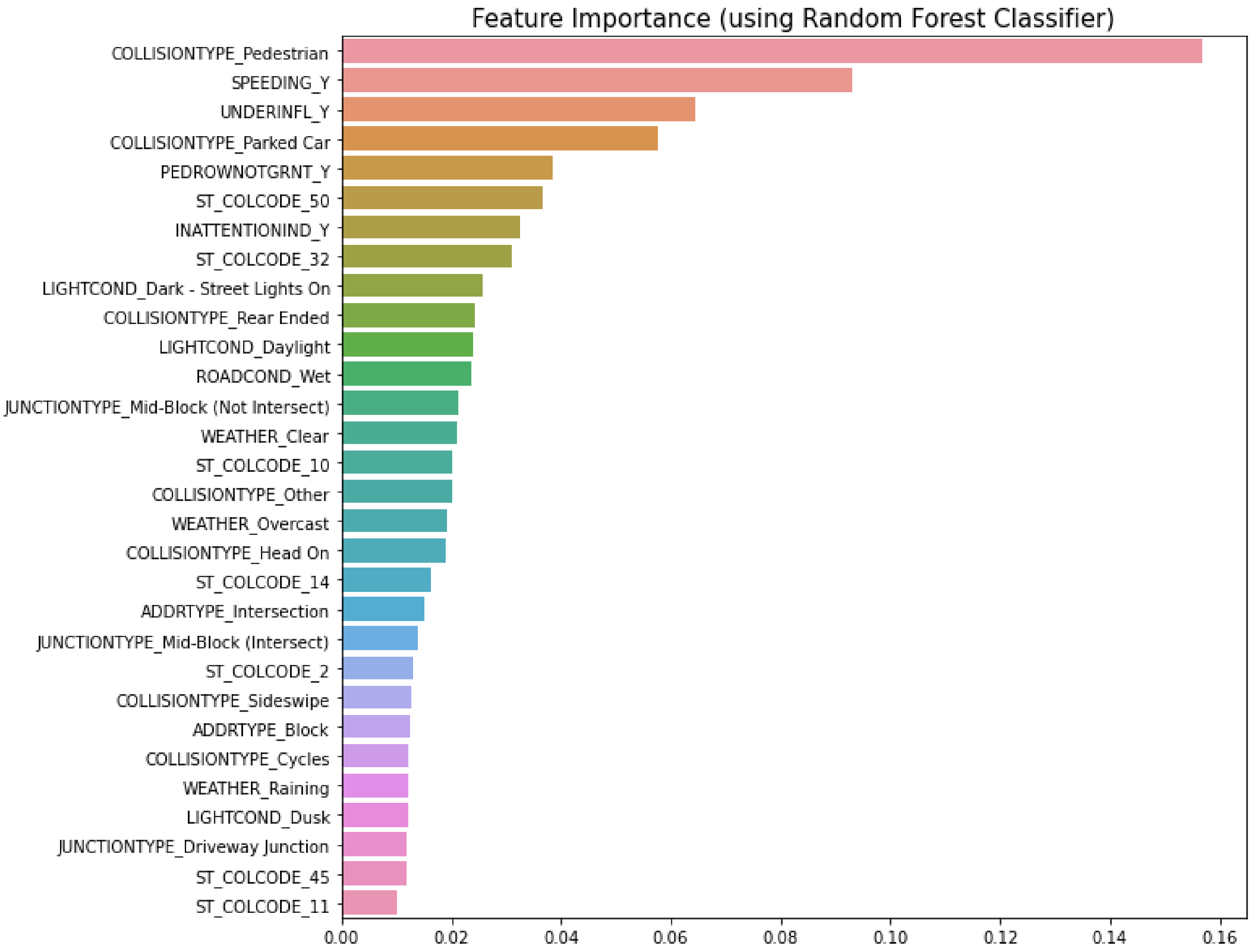


Feature Importance

Decision Tree



Random Forest



Inference

We can conclude that the **Random Forest** is the best model in this scenario, with **Decision Tree** and the other models are almost the same). An interesting point to note here is that the top important features are somewhat different between the **Random Forest** and the **Decision Tree** model. Following the **Random Forest** model, we see that special attention needs to be given to pedestrians (topmost important feature), speeding, collision with a parked car, rear-ended collision, drivers under influence of alcohol/drug. This result is very much expected. The collision codes 50 (Struck Fixed Object), 32 (One Parked – One Moving), 10 (Entering At Angle) and 14 (From Same Direction – Both Going Straight – One Stopped – Rear End) are the major influencers.

Future Study

- The relations between the key features and accident severity can be further studied in details
- Different data balancing techniques can be applied and evaluated
- Development of a much more complex real-time accident risk prediction model