`

Introduction/Business Problem:

The problem we are trying to address is of car accidents in Seattle city. Accidents happen at all times, but if the main causes of accidents are determined, advance warning or mitigating methods can be performed. For example, certain intersections may be more susceptible to accidents due to heavy usage or the way they are constructed. As a result, better street lights can be added (only protected left and right turns) or traffic personnel can be used to direct the cars. We want to analyze the accident "severity" in terms of human fatality, traffic delay, property damage, or any other type of accident bad impact so that advance warning and mitigation strategies can be developed based on insights of what avoidable reasons case these accidents.

The target audience of this analysis is the Seattle government and transportation department. It should identify key causes of accidents and allow them to identify trends for when accidents can be prevented. This will reduce the number of accidents and injuries for the city.

Data:

The data comes from collision and accident reports in Seattle during the years 2004-present. It was collected by the Seattle Police Department and Traffic Records department. There are 194,673 observations and 38 variables in this data set. Since we would like to identify the factors that cause the accident and the level of severity, we will use SEVERITYCODE as our dependent variable Y, and try different combinations of independent variables X to get the result. Since the observations are quite large, we may need to filter out the missing value and delete the unrelated columns first. Then we can select the factor which may have more impact on the accidents, such as address type, weather, road condition, and light condition.
The target Data to be predicted under (SEVERITYCODE 1-prop damage 2-injury) label.

Other important variables include:

- *ADDRTYPE: Collision address type: Alley, Block, Intersection*
- *LOCATION: Description of the general location of the collision*
- *PERSONCOUNT: The total number of people involved in the collision helps identify severity involved*
- *PEDCOUNT: The number of pedestrians involved in the collision helps identify severity involved*
- *PEDCYLCOUNT: The number of bicycles involved in the collision helps identify severity involved*
- *VEHCOUNT: The number of vehicles involved in the collision identify severity involved*
- *JUNCTIONTYPE: Category of junction at which collision took place helps identify where most collisions occur*

• 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)

• 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

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PEDROWNOTGRNT	SDOTCOLNUM	SPEEDING	ST_COLCODE	ST_COLDESC	SEGLANEKEY	CROSSWALK
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight	NaN	NaN	NaN	10	Entering at angle	0	
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On	NaN	6354039.0	NaN	11	From same direction - both going straight - bo...	0	
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight	NaN	4323031.0	NaN	32	One parked--one moving	0	
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight	NaN	NaN	NaN	23	From same direction - all others	0	
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight	NaN	4028032.0	NaN	10	Entering at angle	0	

5 rows × 38 columns

Data Processing:

Many of the observations including the features described above has incomplete information, such as ‘NaN’ (Not a Number) values or bad formatted ones. At the same time, the frequency of the property damage accidents are almost as double as the ones involving injuries. Remember also, that the target variable, the one which will be predicted is SEVERITYCODE.

The data cleaning process must also involve balancing of the data, in this way, the number of entries corresponding to the 2 severities present in the dataset are equal. Severity labeled as 1 correspond to collisions which implies only property damage and the ones labeled as 2, represents personal injuries. This classification is based in the SDOT Traffic Management Division criteria. Let us now explore the data and get the frequency of each category for a given feature. We get the following details:

SEVERITYCODE
1 136485
2 58188
Name: SEVERITYCODE, dtype: int64

ADDRTYPE
Block 126926
Intersection 65070
Alley 751
Name: ADDRTYPE, dtype: int64

COLLISIONTYPE
Parked Car 47987
Angles 34674

Rear Ended 34090
Other 23703
Sideswipe 18609
Left Turn 13703
Pedestrian 6608
Cycles 5415
Right Turn 2956
Head On 2024
Name: COLLISIONTYPE, dtype: int64

JUNCTIONTYPE
Mid-Block (not related to intersection) 89800
At Intersection (intersection related) 62810
Mid-Block (but intersection related) 22790
Driveway Junction 10671
At Intersection (but not related to intersection) 2098
Ramp Junction 166
Unknown 9
Name: JUNCTIONTYPE, dtype: int64

WEATHER
Clear 111135
Raining 33145
Overcast 27714
Unknown 15091
Snowing 907
Other 832
Fog/Smog/Smoke 569
Sleet/Hail/Freezing Rain 113
Blowing Sand/Dirt 56
Severe Crosswind 25
Partly Cloudy 5
Name: WEATHER, dtype: int64

ROADCOND
Dry 124510
Wet 47474
Unknown 15078
Ice 1209
Snow/Slush 1004
Other 132
Standing Water 115
Sand/Mud/Dirt 75
Oil 64
Name: ROADCOND, dtype: int64

LIGHTCOND
Daylight 116137
Dark - Street Lights On 48507
Unknown 13473
Dusk 5902
Dawn 2502
Dark - No Street Lights 1537
Dark - Street Lights Off 1199
Other 235
Dark - Unknown Lighting 11
Name: LIGHTCOND, dtype: int64

SPEEDING
Y 9333
Name: SPEEDING, dtype: int64

UNDERINFL
N 100274
0 80394
Y 5126
1 3995
Name: UNDERINFL, dtype: int64

INATTENTIONIND
Y 29805
Name: INATTENTIONIND, dtype: int64

Cleaning the Dataset: Some of the categories are not relevant or doesn't provide enough information, such as 'Unknown' or 'Other'. We should drop this kind of entries.

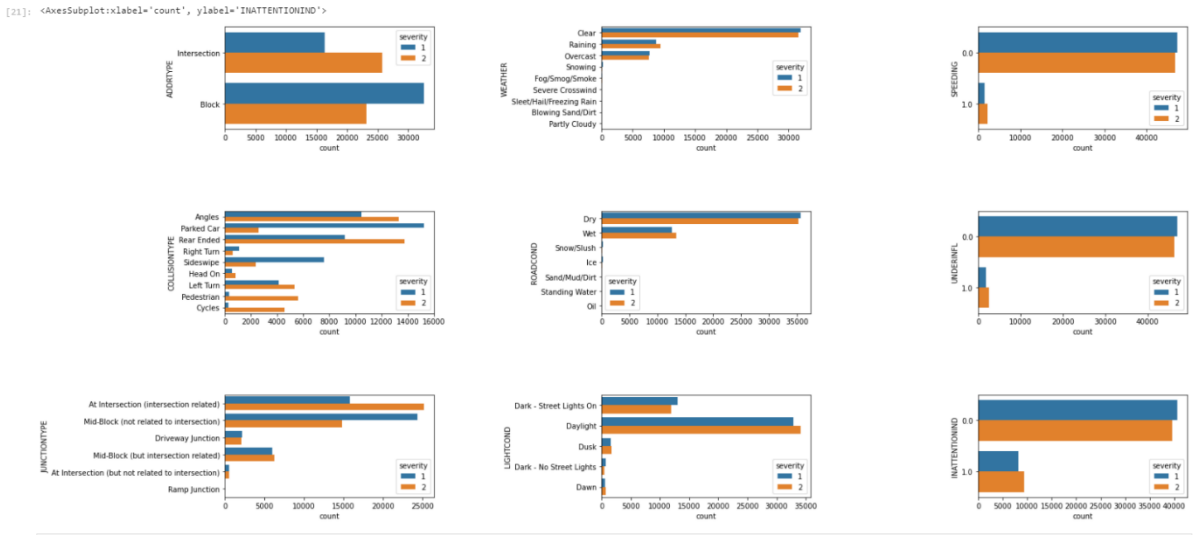
Filling missing data: We assume that if the report does not explicitly indicates an accident which involved driving under a alcohol/substance, then the opposite is true. We will treat the following columns as true only if the field is filled.

Dropping partial entries: Not all the entries or observables are complete, some of them miss relevant information which is needed to train the model.

Post cleaning the data We have a total of 143741 observables. However, there's still some work to do before analyzing and training the model.

Now we need to Balance Data as the not severe collisions, are most frequent than the severe ones. Data is unbalanced, so we will proceed to under sample it.

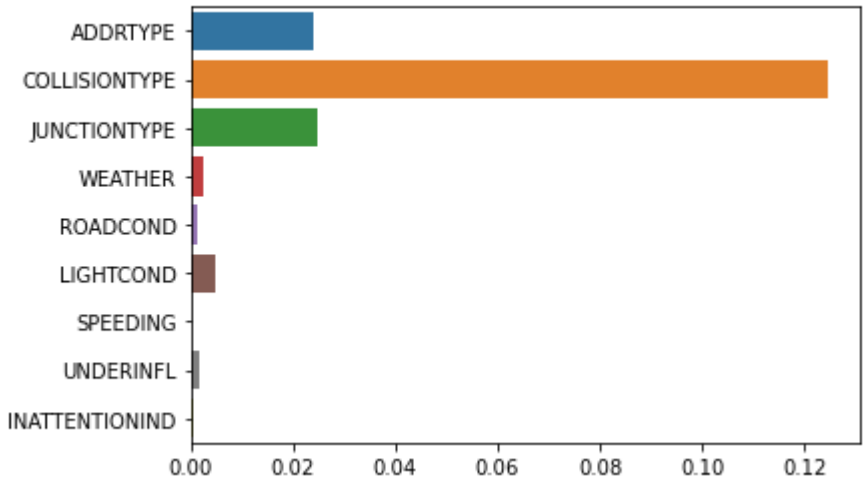
After balancing, we can take a visual look to the variables to see if they are relevant to the model training.



Feature Selection

One of the most important questions before training the model is, are all the features adding the same information to the model? If not so, what variables have more weight on it? To tackle this question some techniques can be used to help select the important features, the ones adding more information to our model. It has to be taken into mind that categorical inputs and output will be used, hence, for this kind of variables there are two common strategies: Chi-Squared Feature Selection and Mutual Information Feature Selection.
Mutual Information Feature Selection will be used for this project. Mutual information from the field of information theory is the application of information gain (typically used in the construction of decision trees) to feature selection. It determines the information entropy for a given variable in a similar fashion as decision trees generates new branches. It is calculated between two variables and measures the reduction in uncertainty for one variable given a known value of the other variable.
Since the algorithm involves splitting the dataset in train and test data in a random fashion, the process was applied 10 times in order to have a smoother information gain for each feature. The result is depicted in the next picture.

[22]: <AxesSubplot:>



It can be observed from the mutual information results, that the variables with more importance in determining the collision severity are: ADDRTYPE, COLLISIONTYPE and JUNCTIONTYPE. There is a clear winner in the result, the COLLISIONTYPE feature. Apparently, the severity of the collision depends noticeably in the location of the car crash (angles, rear side, sideswipe), the maneuvers or whether it involved cycles or pedestrians. Also, the first and the third feature are related, hence is reasonable that the addition of information of these two variables is almost the same.

Feature	Mutual Information importance
ADDRTYPE	0.019883
COLLISIONTYPE	0.126144
JUNCTIONTYPE	0.025643
WEATHER	0.003124
ROADCOND	0.002039
LIGHTCOND	0.004593
SPEEDING	0.000739
UNDERINFL	0.001181
INATTENTIONIND	0.001129

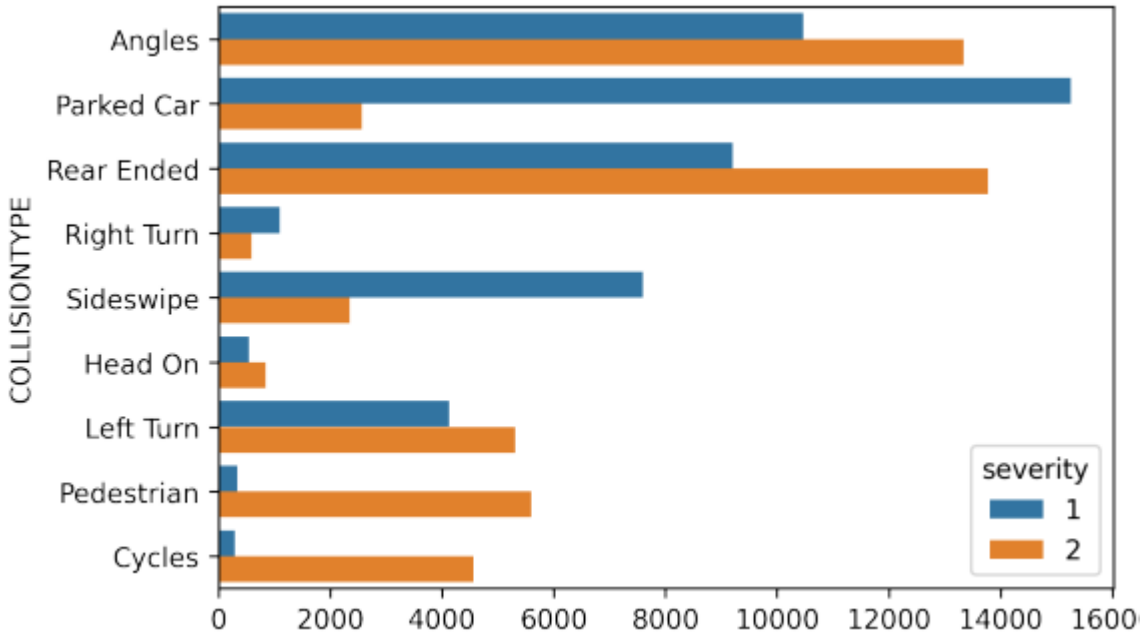
Unfortunately, the categorical variables recently described can not predict the severity of the accident because they are determined after the incident happened. Can we determine beforehand if we will collide with a pedestrian or any kind of cycles? Will it be in an intersection or in the middle of the block? and finally, will it involve a parked car or one in movement?. It is impossible to control these aspects, and for so, they are going be removed from the predictive model.
WEATHER, ROADCOND and LIGHTCOND throws some information to the severity of the accident, this variables can be known before a road user decides to start a journey and will be take into account. Unfortunately there is not too much information provided by this categorical variables.
Lastly, SPEEDING, UNDERINFL and INATTENTIONIND do not add too much information to our target variable, probably as we will see in the next section, because it is not so frequent. This is good news, since nobody wants to deal with fast and furious drivers, under alcohol or drugs influenced ones or inattention conductors, such as the ones texting with their mobile phones.

Exploratory Analysis

Before introducing our data to the classifier algorithms, let's explore the data to see if we can gather some knowledge from it and get some insights. It is also important to have in mind that some variables chosen can not be used to create a predictive model, since it is based in information collected after the accident had taken place. The data analyzed in the following paragraphs has balanced events for each severity. Impacts that implies only property damage are labeled as 1, which are almost as double as frequent than severe ones. These type 1 labeled impacts, have been under-sampled. For further information please visit the Data Cleaning subsection.

Severity and Collisions

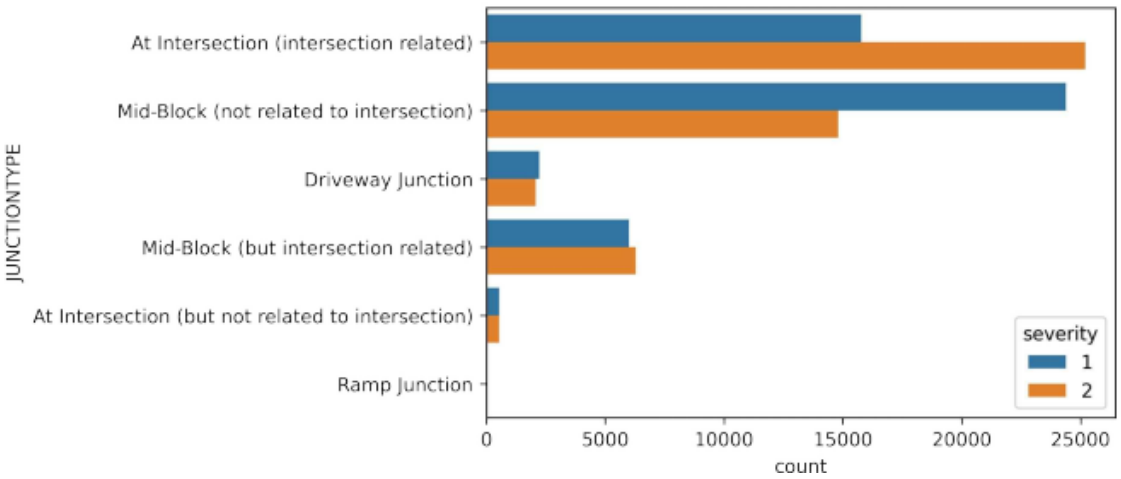
One important aspect revealed by the data is related to the severity of accidents based on the collision type. This feature has different characteristics based in the area of impact, such as: angles, parked car, rear end, right turn, sideswipe, head on, left turn, pedestrian and cycles. All those variables, their frequency and the collision severity can be found in the figure below. As can be observed, the entropy of this categorical variable is pretty high (very unbalanced).



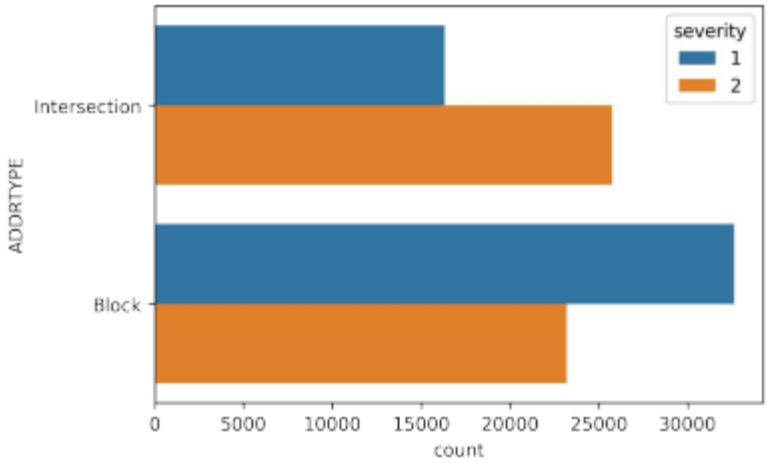
Accidents involving car damaged parts such as angles or the rear side as well as left turns tend to be more dangerous, since involves some personal damage to the car users. Also, those collisions involving pedestrian or cycles in general are riskier than others. Furthermore, it is important to notice that left turns are generally riskier than right ones, which might be related to those turns from avenues to streets. The scenarios described in the paragraphs above involve the worst case scenarios, including personal injuries. However, some other impacts in different scenarios such as car parking or sideswipes - generally speaking- involve less risk to car passengers.

Mid-Block vs. Intersections

Looking at the following histogram, we can observe a higher frequency for severe accidents in intersections rather than in the middle of the blocks. In the same way, mid-block collisions are not severe, involving only car damage in most of the cases.

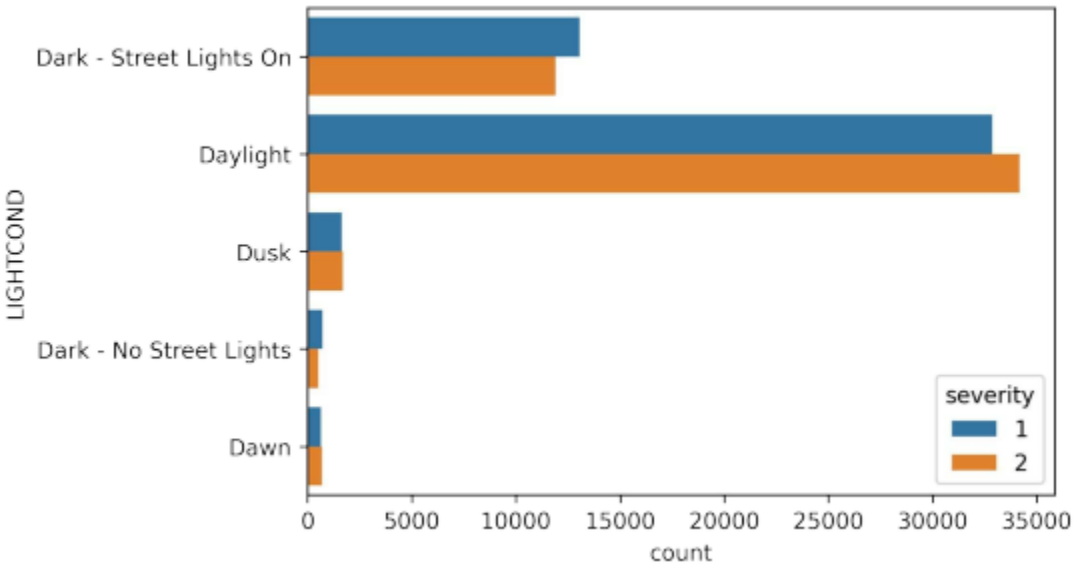


There are some less frequent categories which can be grouped in a meta-category involving intersections and mid-block incidents. Since these categories are relatively balanced, the overall classification does not change. Indeed, the categorical variable ADDRTYPE divides collision in these two categories, as we commented in the Feature Selection subsection. Not surprisingly, the amount of information which JUNCTIONTYPE and ADDRTYPE adds to the model is similar. This analysis can be inferred from the picture below.

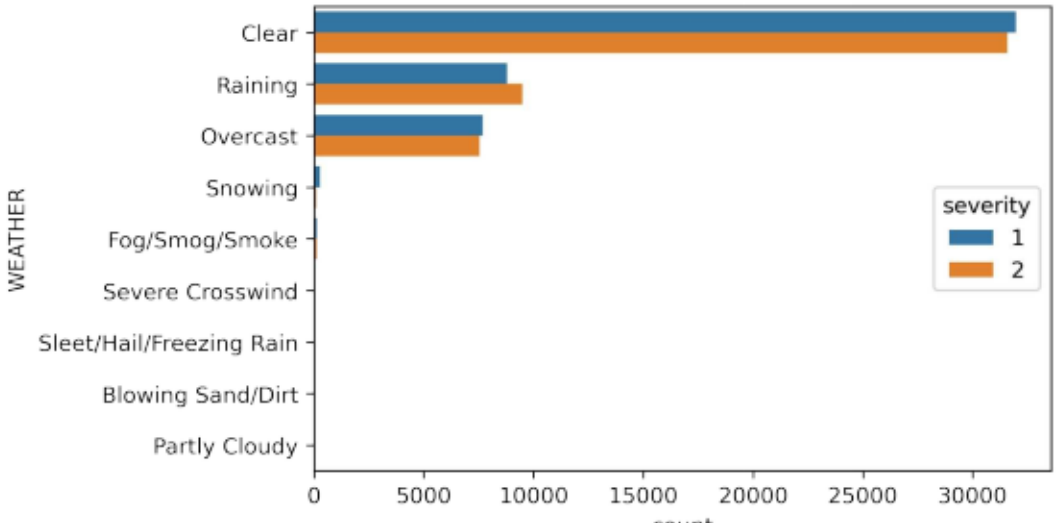


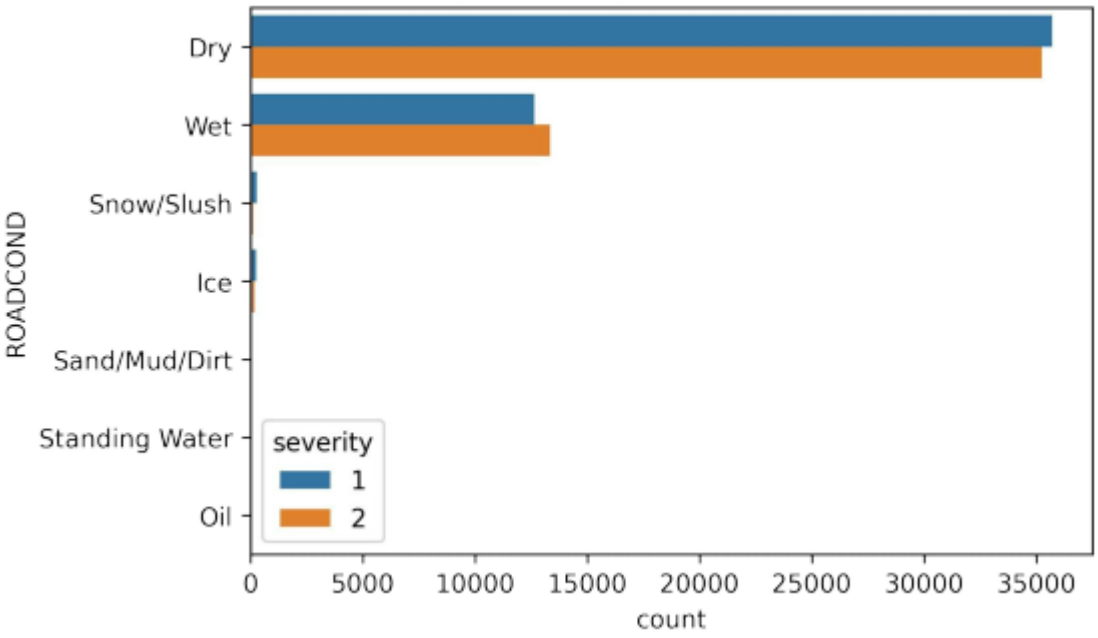
Weather, Light and Road conditions

Weather and light conditions reveal some insights, although the differences in the severity is only slightly biased to one of the sides. There are more occurrences of severe collisions during daylight whereas car drivers tend to have less injuries while driving during the night over streets with the lights on. The reason for this, may be related to a more cautious driving during the night which predispose the people to a state of awareness. Dusk and dawn tend to be related to more severe collisions, maybe because of the visibility reduction while facing the sun directly in the vision zone

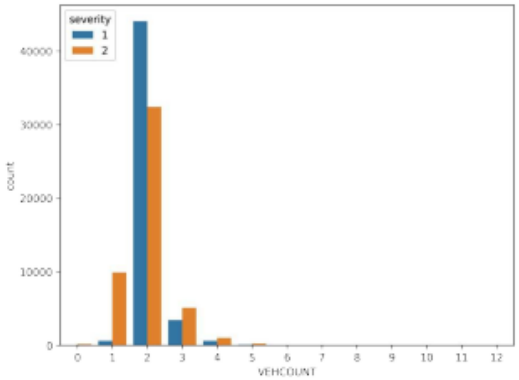


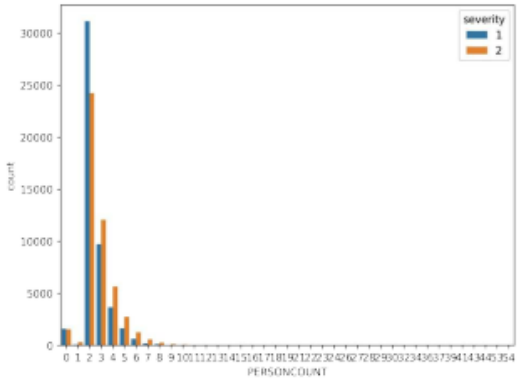
Considering weather data, severe accidents are slightly more frequent during rainy weather as well as with wet roads. However data is pretty balanced between both severity types and for this reason, the lack of entropy do not add too much information to our model.





Severity based on number of cars and people involved
Collisions involving multiple car or people tend to be more severe than others. However, at the same time they are more infrequent. The plots reveal also, that most of the times accidents involve 2 people and 2 vehicles. The impacts tend to be not so severe, implying -luckily- most of the times, damage to the participants cars.





Unfortunately, this is information which can not be taken beforehand, since we can not anticipate to the occupants number and the vehicles involved in a collision. However the data adds knowledge of the overall collisions behavior when the number of vehicles and people involved, is taken into consideration.

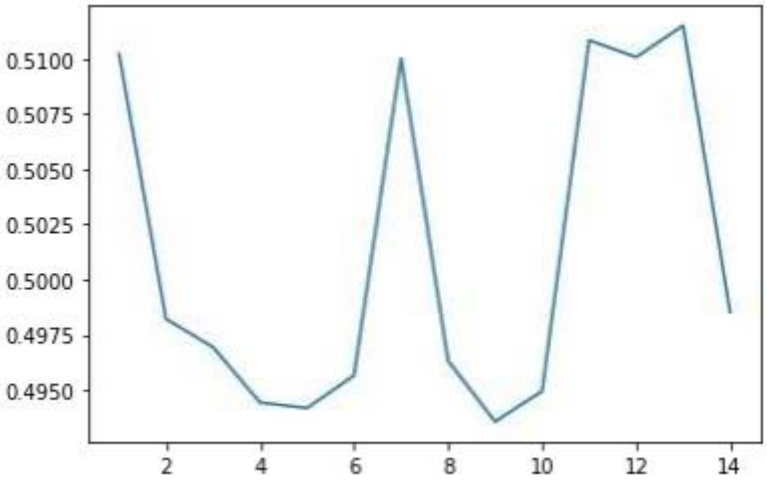
Predictive Modeling

Given the data provided in the database and the target variable, it is clear that we are under a problem involving categorical inputs and outputs. Since this is the scenario, the type of machine learning algorithms to apply are classification ones. We are going to use in particular the following ones: KNearest Neighbor (KNN), Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF). The model with the best results will be optimized in order to fine tune it and compare the difference with the standard values.
One of the things we have to do before training the model is to relabel our target variable as 0 for collisions with severity involving only property damage and 1 for those involving injuries. After that, it is important to select variables which can perform the best in the classification process and at the same time can have predictive use. For this reason and for what was commented in the feature selection paragraph, the predictors are WEATHER, ROADCOND and LIGHTCOND.
Another important thing to take into account is how to encode the categories in each predictor, since each variable has several categories. The best approach is to convert everything to one hot encoding to avoid the model getting lost in hierarchy issues present in label encoding methods. In those cases, the model may try to predict values which are in the middle of a category. However, this is not representative of the observable universe, because they do not actually exist. Finally, to avoid biases, it is important to normalize information before entering our observation matrix to the training.

Results

KNN Modeling

The first thing to do when treating with this algorithm is finding the optimum k parameters to later train the model. To do so, it can be iterated over a set of k values in a given range and find the one which produces the less error when the predictions are compared to the real values. For this case, it have been relied in the accuracy score to select the best k.



Although higher k values may provide a best estimation, there is always a risk of over fitting the model. In this way, the performance will be above average for the given set of observables but not as efficient for predicting new entries. It has been decided to choose a k value of 7.
The classification report for this model in particular is

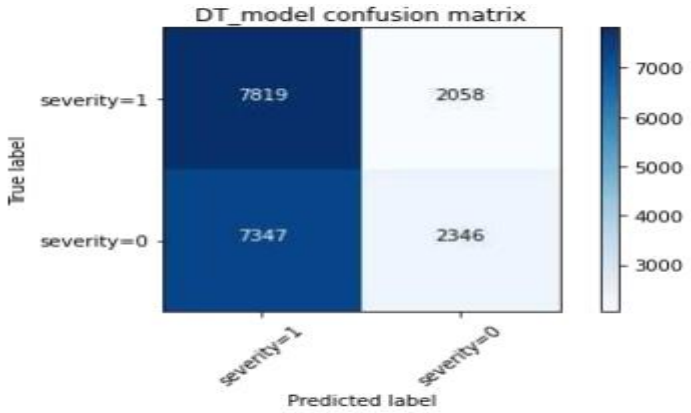
	precision	recall	f1-score	support
0 – property damage	0.51	0.43	0.46	9693
1 – personal injuries	0.51	0.59	0.55	9877
accuracy			0.51	19570
macro avg	0.51	0.51	0.51	19570
weighted avg	0.51	0.51	0.51	19570

Decision Tree Modeling

Besides the default parameters, we have indicated to use the entropy criterion to create the branches, which goes in line with the idea followed when the feature selection was performed. In this way, the strategy follows the same criteria, opening branches that generates the most information gain. We have not specified a limit to the tree depth nor any other parameter. The classification report can be found in the following table.

	precision	recall	f1-score	support
0 – property damage	0.53	0.24	0.33	9693
1 – personal injuries	0.52	0.79	0.62	9877
accuracy			0.52	19570
macro avg	0.52	0.52	0.48	19570
weighted avg	0.52	0.52	0.48	19570

We also include the correspondent confusion matrix of the model in the following image

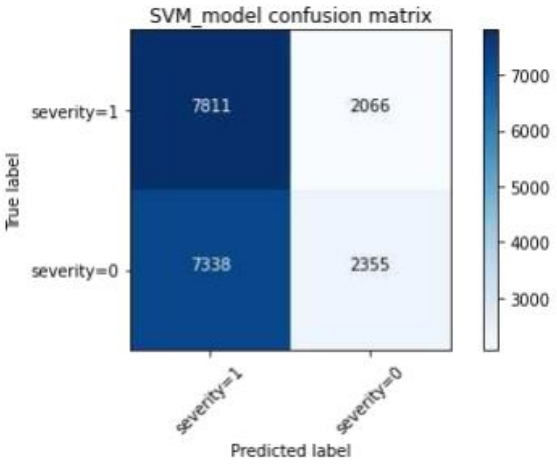


Support Vector Machine Modeling

For this particular model we have left all the parameters by default. This involves using the following criteria: regularization parameter C=1, kernel='rbf', gamma=scale. Those 3 parameters are the most fined tuned ones. KNN and SVM algorithms are good for small samples or datasets but not too much effective with large ones, since they are very intensive in computational terms. The classification report for this algorithm is as follows:

	precision	recall	f1-score	support
0 – property damage	0.53	0.24	0.33	9693
1 – personal injuries	0.52	0.79	0.62	9877
accuracy			0.52	19570
macro avg	0.52	0.52	0.48	19570
weighted avg	0.52	0.52	0.48	19570

Notice that the results are the same as the ones obtained with the decision tree model, however this is just a coincidence and not a report error. Below you can find the confusion matrix results.

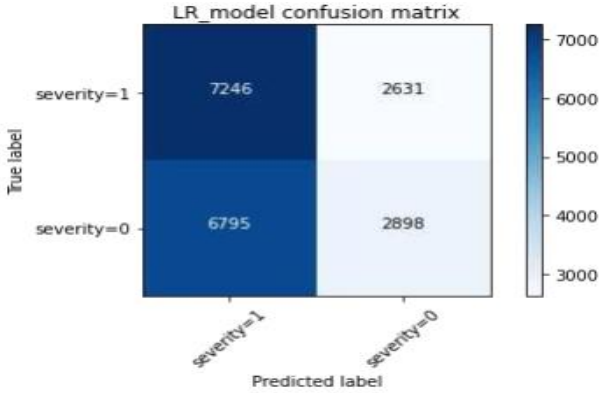


Logistic Regression Modeling

For the LR model, we have chosen the inverse of regularization strength in C = 0.1. The other parameters were left as default. This selection of C value aims to develop a model which is less prone to over-fitting. The smaller the number, the less the chance of over-fitting. The results of the classification report can be found in the following table.

	precision	recall	f1-score	support
0 – property damage	0.52	0.30	0.38	9693
1 – personal injuries	0.52	0.73	0.61	9877
accuracy			0.52	19570
macro avg	0.52	0.52	0.49	19570
weighted avg	0.52	0.52	0.49	19570

Below you can find the results of the confusion matrix for this particular classifier.

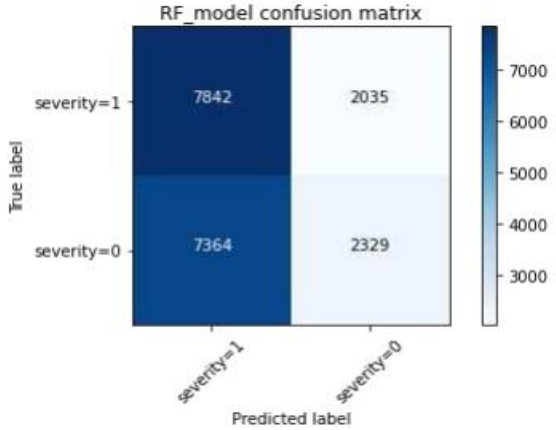


Random Forest Modeling

The Random Forest algorithm introduce randomness in the classifier construction when compared with traditional Decision Tree ones. The prediction of the ensemble is given as the averaged prediction of the individual classifiers.
The purpose in generating randomness is to decrease the variance of the forest estimator. Indeed, individual decision trees typically exhibit high variance and tend to overfit. The injected randomness in forests yield decision trees with somewhat decoupled prediction errors. By taking an average of those predictions, some errors can cancel out. Random forests achieve a reduced variance by combining diverse trees, sometimes at the cost of a slight increase in bias. In practice the variance reduction is often significant hence yielding an overall better model.
The classifier was trained with default parameters, the summary of the classifier can be seen in the following report table.

	precision	recall	f1-score	support
0 – property damage	0.53	0.24	0.33	9693
1 – personal injuries	0.52	0.79	0.63	9877
accuracy			0.52	19570
macro avg	0.52	0.52	0.48	19570
weighted avg	0.52	0.52	0.48	19570

The performance is similar as the Decision Tree algorithm. Below you can find the confusion matrix result.

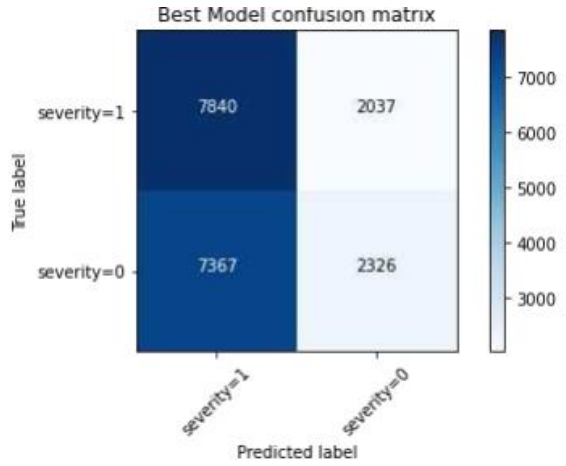


Optimization

Analyzing the model results, we can see that DT, SVM, LR and RF performs in a similar fashion. Since DT and RF has some similar features, we are prone to optimize the later to see if we can fine tune the results. Between the others, taking into account that KNN and SVM are not optimal for large datasets due to their intensive processing power use and similar results, LR was chosen as the other model to optimize.
The best performing model between different parameter values is RF with the following values: max_features=6, n_estimators=75.
The classification report is summarized in the following table.

	precision	recall	f1-score	support
0 – property damage	0.53	0.24	0.33	9693
1 – personal injuries	0.52	0.79	0.63	9877
accuracy			0.52	19570
macro avg	0.52	0.52	0.48	19570
weighted avg	0.52	0.52	0.48	19570

Below is depicted the confusion matrix for the optimized Random Tree classifier. As it can be noticed, accuracy and recall is similar to the other models, there are only slight changes



Summary

Based on the results obtained after training the different models, they are displayed in the table below. Briefly, the training outcomes shows that there is no model which outperforms the rest.

model	accuracy	jacccard	f1
KNN	0.51	0.38	0.51
Decision Tree	0.52	0.45	0.48
SVM	0.52	0.45	0.48
Logistic Regression	0.52	0.43	0.49
Random Forest	0.52	0.45	0.48
Random Forest (opt.)	0.52	0.45	0.48

Accuracy is roughly the same for each model. That is, the ability to predict correctly given the set of predictors is about 50% which is the same as flipping an unloaded coin in the air.
The jaccard index which measures the similarity between the test and predicted set and ranges from 0 to 1, give us in the best approximation 0.45. This means that there is more than half of the test set left out of the predicted model.
Finally, the f1-score which is determined based on the precision and the recall average of the sets, while predicting both target scenarios throws a similar number as the other scores, without showing too much dissimilarity between each prediction model.
We selected before, the two best models to perform a fine-tuning of the parameters, however the results does not differ significantly from the default values. There was not too much gain in the process.

Discussion

Much of the information recollected from the database has been analyzed in different plots. The results of them threw significant information. In the data analysis process the focus was to understand which were the factors that can predict the accident type better.
Collisions which does not involve personal injuries are twice as frequent as the ones involving lesioned ones. Much of the useful information for classification is embedded in the post-collision data collected. From this information, it was learned that collisions involving cycles or pedestrians are severe and involves injuries. It is important to take care of them since they are highly vulnerable to traffic incidents. Some efforts are being held to mitigate those risks providing interurban trails, to transit with bikes or simply walking. The downtown, which has the highest collision record, have implemented many protected bike lanes, and multi-use lanes, shared with pedestrians, that extends to nearby neighborhoods such as Queen Anne, Capitol Hill, between others.
The riskier car collisions are the ones that hit the car from the rear end. This characteristic can aid car automakers to improve the vehicles design in order to mitigate the effects of this kind of collisions. At the same time the frequency of this type of collision is quite high, what put them in central debate. Some studies affirm that many of these accidents are caused by distracted drivers, fatigue, aggressive (speeding) and drunk driving. Efforts are being held to mitigate this incidents with the implementation of crash avoidance systems which take the car brakes control if there is a risk of collision with the car in front of the first. It is rather important and highly recommended, that car users choose this as a safety feature.
Safety at unsignalized intersections is a major concern. Intersection collisions are one of the most common types of crash, and in the United States, they account for nearly 2 million accidents and 6,700 fatalities every year. However, a fully signalized intersection can sometimes be hard to justify in rural areas, due to the cost of installation, maintenance, and added delays to traffic on the major through streets. The Intersection Collision Warning System (ICWS) project studied the effectiveness of an innovative and potentially less expensive approach to improving safety in these situations. This approach consists of two types of traffic-actuated warning signs linked to pavement loops and a traffic signal controller. Concerning the particular Seattle situation, using the same database, in this page a list of the most dangerous intersections can be found.
The last picture show us that the Downtown and the North-Eastern neighbors are the ones with more events. These neighbors should take more attention and further evaluation from the local government and transportation division to increase infrastructure and to reduce the collision incidences. Clearly this is the hot accident spot in the metropolis where car users have to pay extra attention in their maneuvers.

Conclusion

Much of the data analyzed had revealed, some important information about car accidents. Concerning the riskier ones which involve personal injuries, the focus has to be made in some important factors: intersections, rear end collisions, pedestrian and cycles. Left turns are also risky maneuvers which should also be avoided if the road users want to be safe.
Extremely dangerous weather and road conditions do not produce a quite significant accident rate, such as snow and ice. However ,caution have to be taken with rainy weather and wet roads, since after clear days and dry roads, those are the following conditions in order of importance.
Finally, the results of the machine learning algorithms using predictors such as the weather, road and light conditions throws mediocre results. Other factors have to be considered to improve the prediction rate of the models being used.