Introduction

Car accidents are one of the most undesirable and unexpected events which may happen to a road user. Drivers have to face destiny every day and assume the risk to drive to their final destination loaded with a high amount of uncertainties. There may be several reasons to start a car journey: commuting to work, going out to dinner, visit friends or family or even going out for holidays. No matter the reason, the risk is always embedded in the decision.

In year 2010, there were 32,999 people killed, 3.9 million were injured, and 24 million vehicles were damaged in motor vehicle crashes in the United States. The economic costs of these crashes totaled $242 billion. Included in these losses are lost productivity, medical costs, legal and court costs, emergency service costs (EMS), insurance administration costs, congestion costs, property damage, and workplace losses. This represents a 1.6 percent of the $14.96 trillion real Gross Domestic Product for 2010.

The society as a whole - the accidents victims and their families, their employers, insurance firms, emergency and health care personal and many others - is affected by motor vehicle crashes in many ways. It would be great if we can provide real-time conditions and estimate the trip safeness. In this way, we can decide beforehand if we are willing to take the risk, based on reliable information.

Business Problem

The aim of the project is to use relevant information which can help road users, insurance companies, employers, health care providers, road maintainers and traffic congestion models, between others. Using Data Analysis and training different Machine Learning models, the best one will be chosen to predict the severity of a car accident based in some variables that will work as predictors of the incident.
The model will include variables such as weather, road and light conditions. It will also generate some insights in accident severities based on the amount of people traveling in the car, the areas where accidents take place most frequently, impact of driving under substances abuse and also distraction while driving, such as using the mobile phone and some others. It will also seek to provide a better understanding of the actual road conditions, likeliness of an impact, how to reduce incidents and the hot accident zones.

Target Audience

The results will help road users, insurance companies, employers, health care providers, road maintainers and traffic congestion models, between others in the Seattle, WA. area. It will also help to have a better understanding of the actual road conditions, likeliness of an impact, how to reduce incidents and the hot accident zones. These insights can be helpful to the government sector in how to develop new infrastructure.

Data

The data used in the analysis is provided by the Traffic Records Group in the SDOT Traffic Management Division from Seattle, WA. Their work is centered on a transportation system that provides safe and affordable access to places and opportunities. The data includes all collisions

provided by the Seattle Police Department and recorded by the Traffic Record, displayed at the intersection or mid-block of a segment from 2004 to the present.

The attributes to use are described in the following table

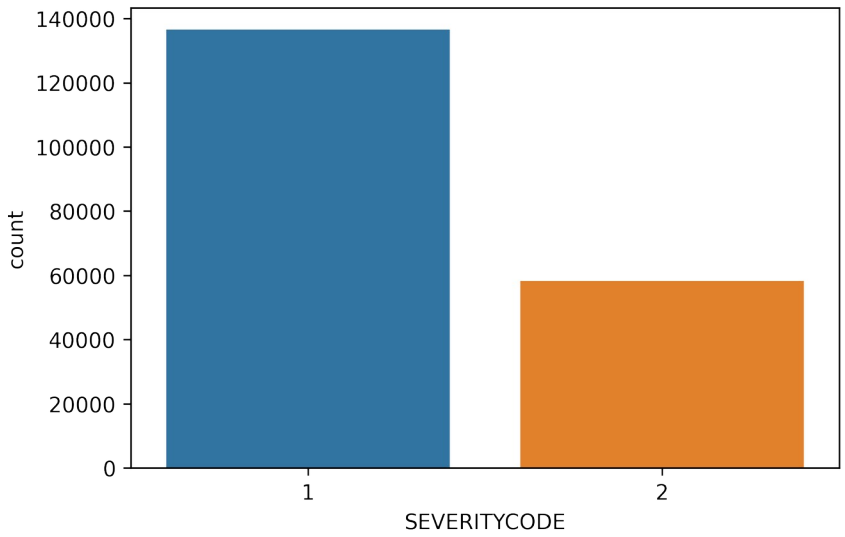
Feature	Description
LOCATION	Latitude and longitude of the incident.
ROADCOND	Status of the road at the moment of collision.
WEATHER	Whether conditions at the moment of collision.
ADDRTYPE	Whether collision occurred in block or intersection.
JUNCTIONTYPE	Detailed description of the place of collision.
COLLISIONTYPE	Type of collision: rear, angles, sideswipe, etc.
SPEEDING	Whether driver was speeding during incident.
LIGHTCOND	The lights condition during the collision.
NUMCOUNT	Number of people involved in the accident.
VEHCOUNT	Number of vehicles involved in the accident.
UNDERINFL	Whether the driver was under alcohol or drugs influence or not.
INATTENTIONIND	Whether or not the collision was caused by inattention.
SEVERITYCODE	A code to describe the severity of the collision.

Features from collision database to be analyzed and used

Our target attribute is the severity of the collision which in the database is categorized as 1 if there is only property damage and 2 if it includes personal injuries. No fatalities were found in the given database, so the classifying method will be based in these two categories.

Data cleaning

Many of the observations including the features described above has incomplete information, such as ‘NaN’ values or bad formatting. At the same time, the frequency of the property damage accidents are almost as double as the ones involving injuries. The data cleaning process must also involve balancing of the data, in this way, the number of entries corresponding to severity 1 and 2 are equal.



Entries classified by severity before cleaning and balancing the data

Some features have categories such as “Unknown” or “Other” which are not representative and do not add predictive information to the training model. These categories together with the empty fields which do not have a valid entry, will also be dropped.

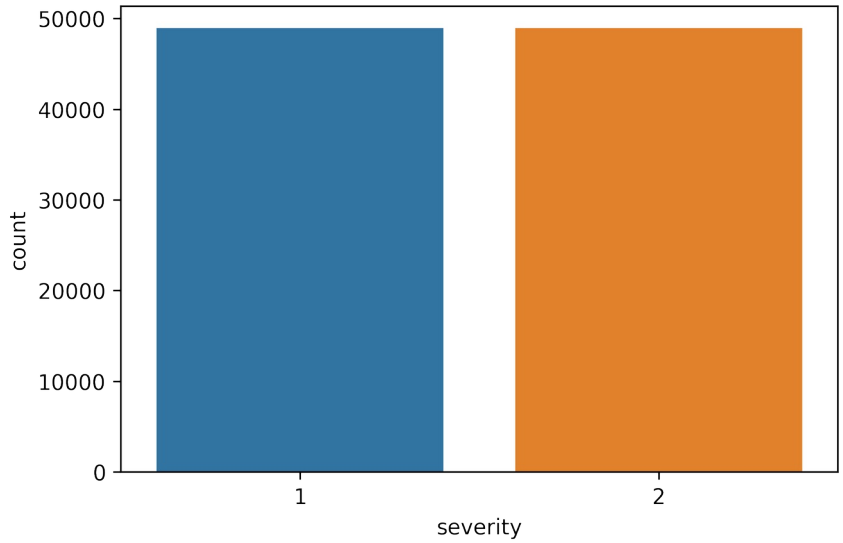
LIGHTCOND feature has some entries described as “Dark - Unknown Lighting” which also be removed and “Dark - Street Lights Off” will be merged with “Dark - No Street Lights”.

The UNDERINFL feature has some entries as 0 and other as N, as well as some others as 1 or Y. The same occurs with INATTENTIONIND and SPEEDING. All of them will be transformed to 0 or 1.

Finally, all the incomplete fields in UNDERINFL, INATTENTIONIND and SPEEDING features, will be completed with 0. This is, given there is no information about these features for the particular collision, innocence is presumed. In this way, we can increase the number of observables for those variables, which in the other case, will only represent a marginal number.

As mentioned before, the data needs to be balanced between the two categories in order to improve the accuracy of the predictive machine learning models (unless decision tree like models are trained). For this purpose, the imbalanced learn library has been used and particularly the RandomUnderSampler object to perform an under-sampling of the dataframe. This strategy, eliminates randomly the extra entries corresponding to severity grade 1. Up to the point where there is the same amount of entries with both severities.

After the data cleaning and under-sampling process, the frequency for each severity is as follows. The details of frequency per each feature, will be discussed in the following sections.



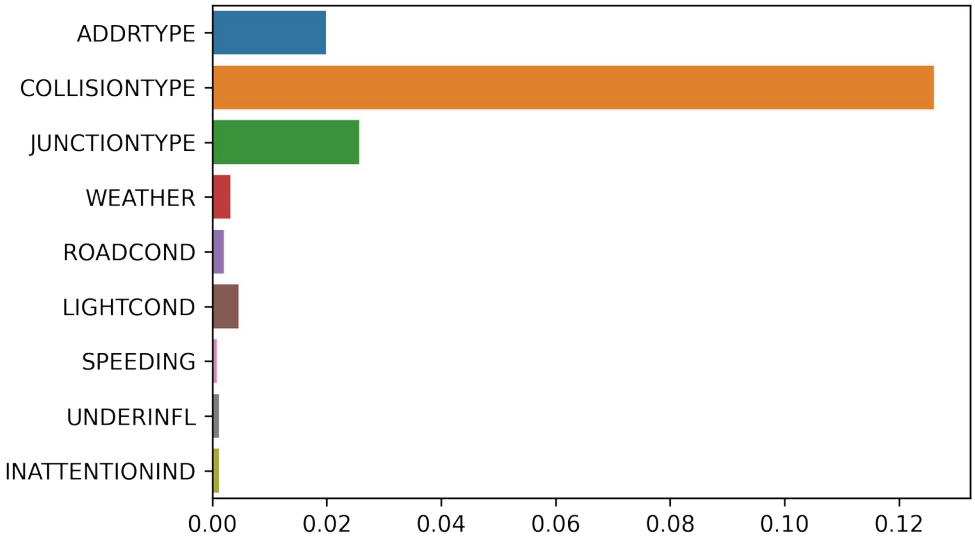
Collisions classified by severity after cleaning and balancing the data

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, how can we determine which variables are influencing more the predictiveness of it? To tackle this question we can use some techniques which can help us to select the important features, the ones adding more information to our model. We have to take in mind that we are dealing with categorical inputs and a categorical output, hence, for this kind of variables there are two common strategies: Chi-Squared Feature Selection and Mutual Information Feature Selection.

We will use Mutual Information Feature Selection for our 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.



Mutual information importance per categorical inputs

We can observe 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

Features and their corresponding mutual information values

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, cycles? If it will be in an intersection or the middle of the block and if it will involve a parked car or one in movement?. It is impossible to control these aspects, and for so, they will 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

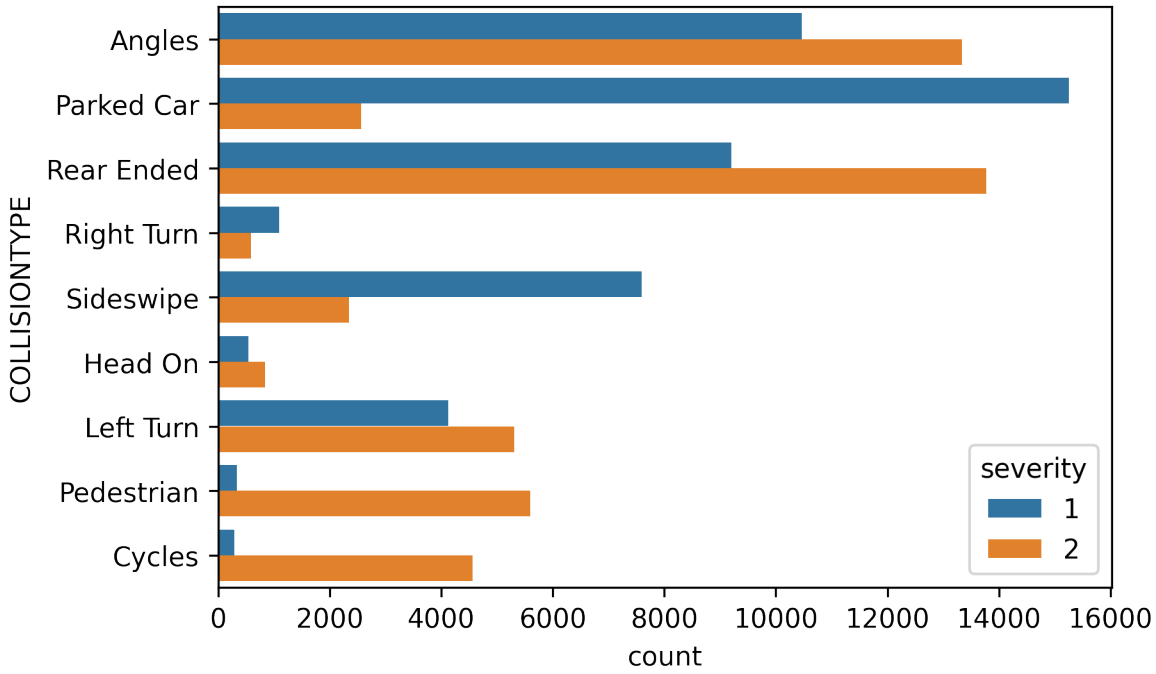
news, since nobody wants to deal with fast and furious drivers, under alcohol or drugs influenced ones or inattention conductors, such as the ones texting with their mobile phones.

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).



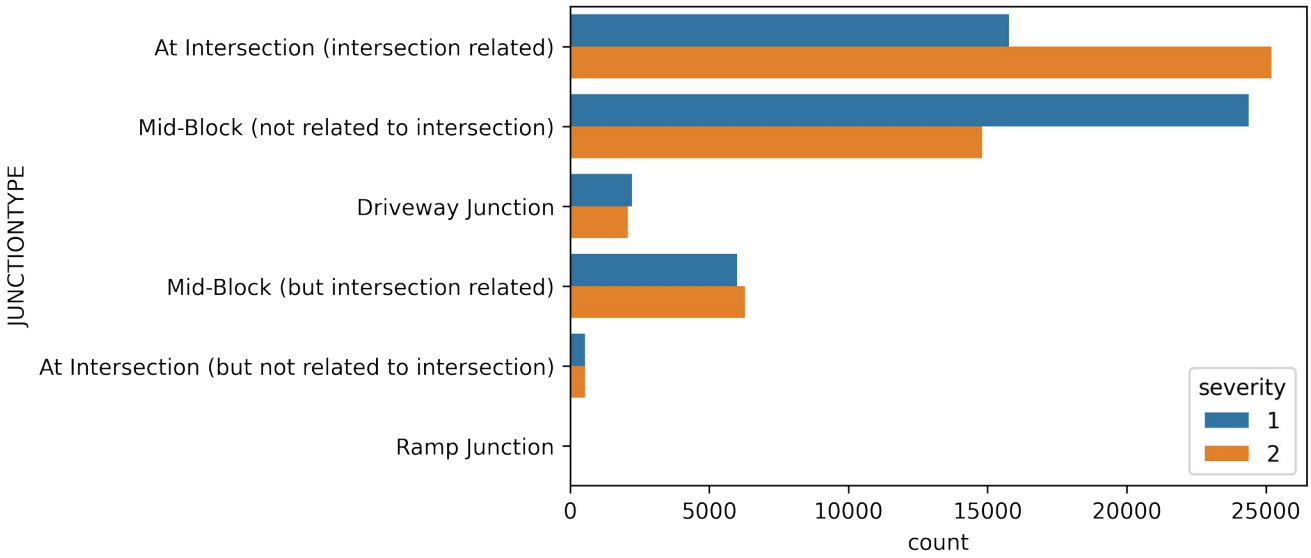
Collision categories, frequency and severity

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.

It has been described some scenarios where there severity is prone to personal injuries. However, some other impacts in different scenarios such as car parking or sideswipes -generally speaking- involve less risk to car passengers.

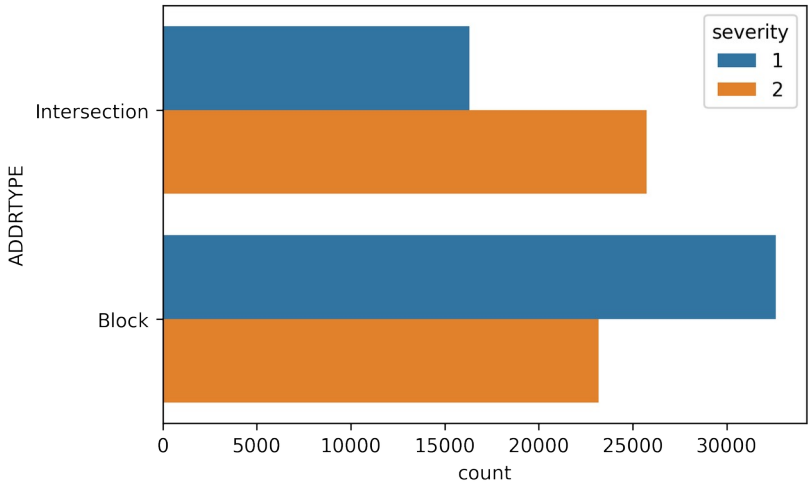
Mid-Block vs. Intersections

Looking at the following histogram, we can observe a higher frequency for severe accidents, which are more common 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.



Frequency per junction type collision and its severity

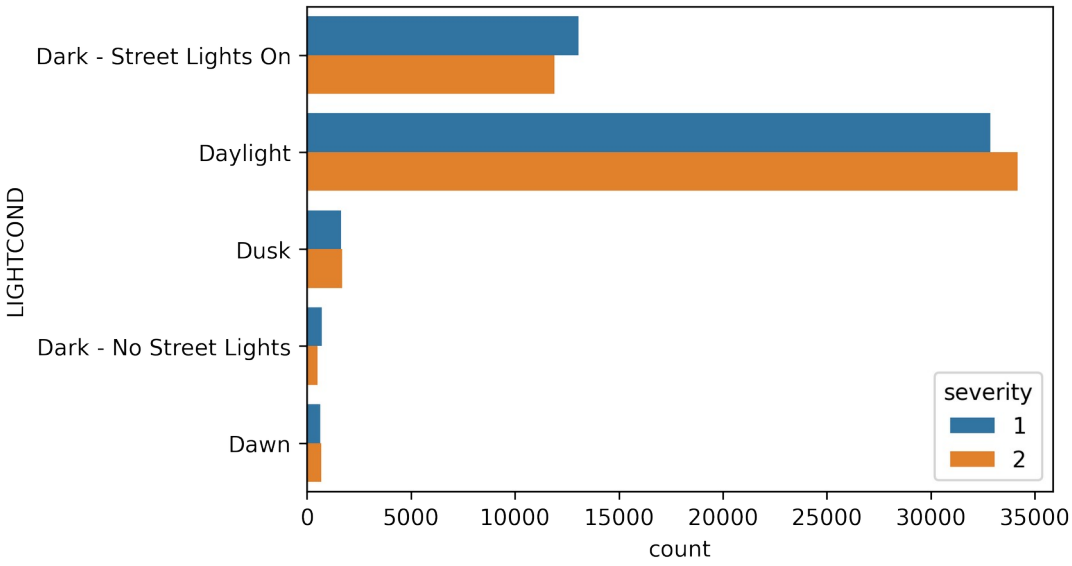
There are some less frequent categories which can be grouped in a meta-category involving intersections and mid-block incidents. Since these categories are relatively balanced, the overall classification does not change. Indeed, the categorical variable ADDRTYPE divides collision in these two categories and 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.



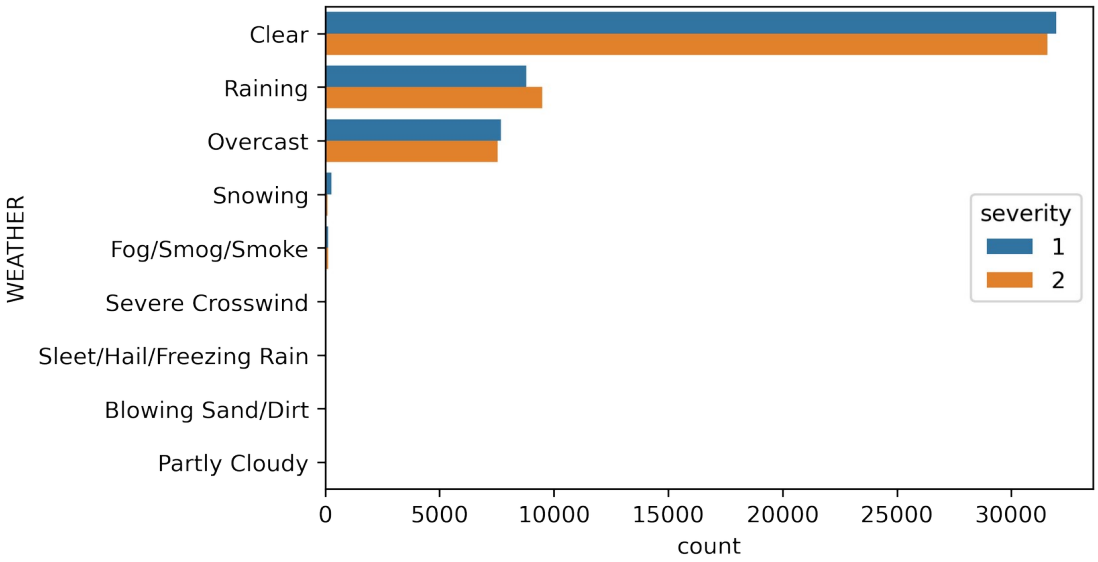
Frequency for each address type collision and its severity

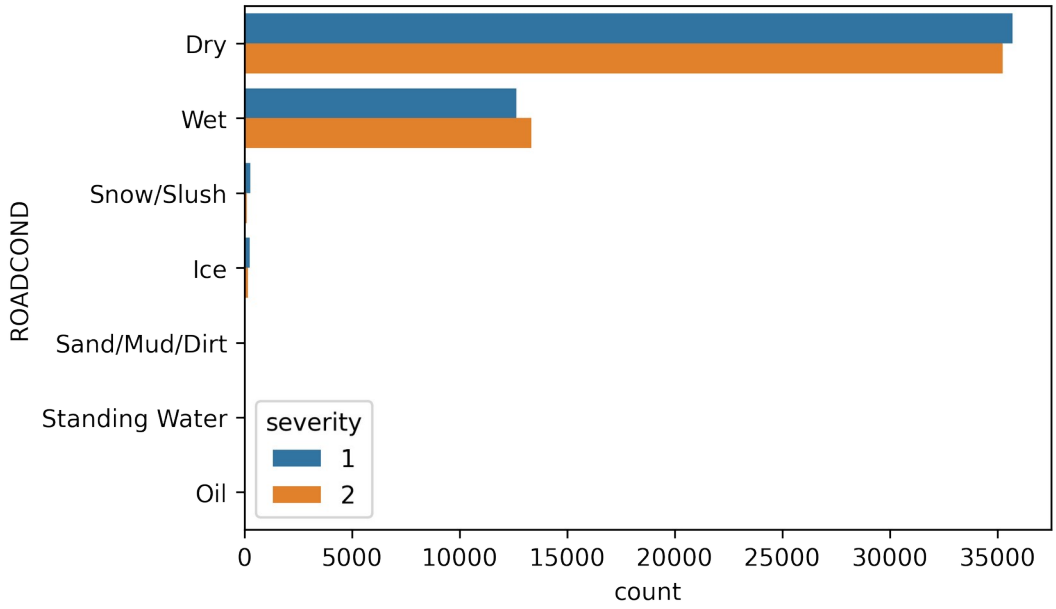
Weather, Light and Road conditions

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.



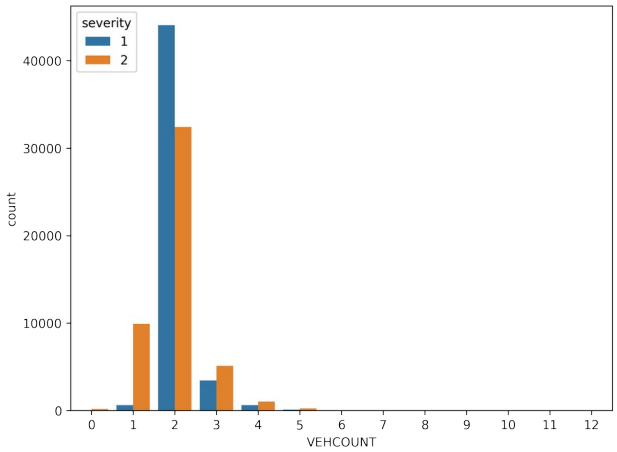
Considering weather data, severe accidents are slightly more frequent during rainy weather as well as with wet roads. However data is pretty balanced between both severity types and for this reason, the lack of entropy do not add too much information to our model.

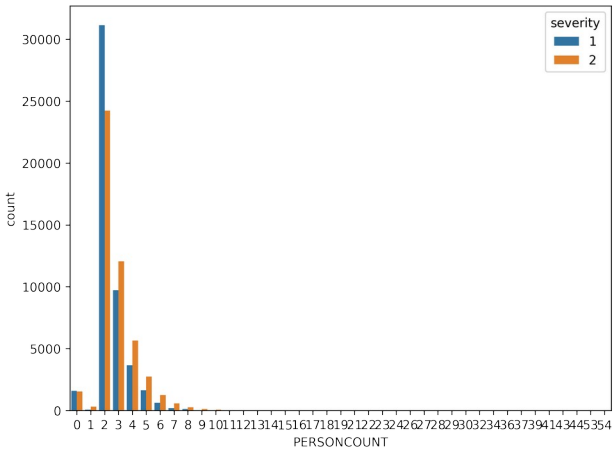




Severity based on number of cars and people involved

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.

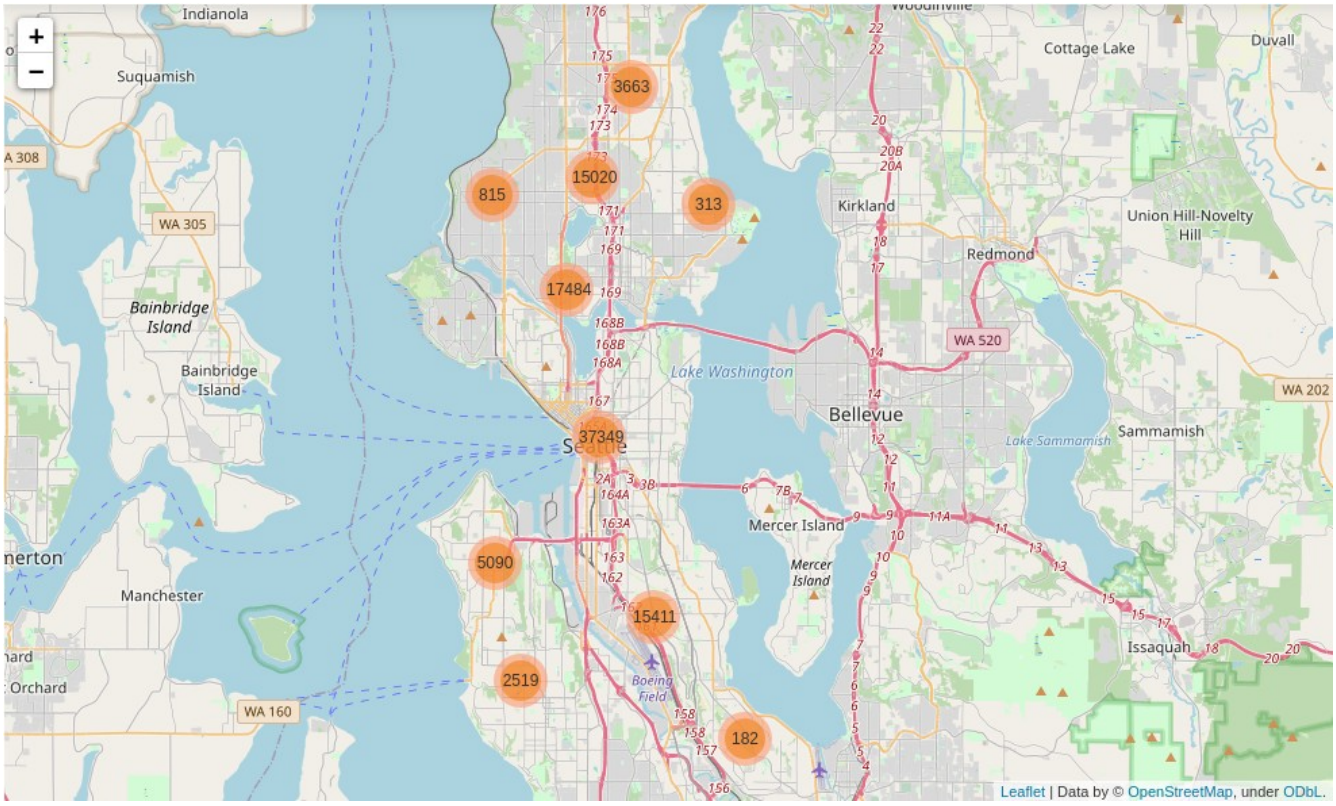




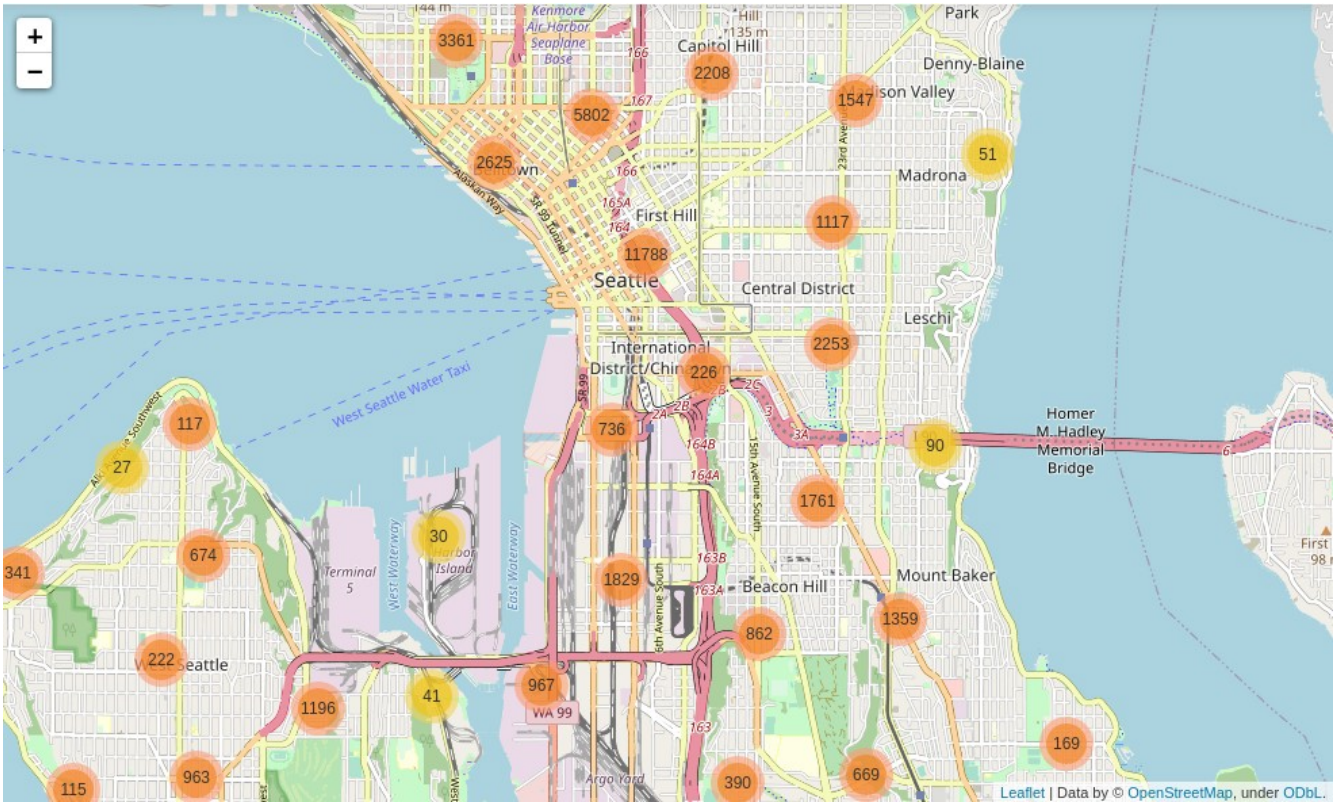
Unfortunately, this is information which can not be taken beforehand, since we can not anticipate to the numbers of occupants and vehicles involved in a collision. However the data adds knowledge of the overall collisions behavior when the number of vehicles and people involved, is taken into consideration.

Clustering collisions in different geographic areas

One relevant question to explore is, where are most of the accidents occurring? Are there some hot zones where is most probable to have collisions?. In the following pictures a map including clusters with incidents is displayed.



It is pretty clear that most of the accidents occurs in Pioneer Square, Yesler Terrace, the Downtown, Belltown and Pike/Pine. A closer picture let us recreate smaller clusters which shows further information about these areas.



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.

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: K-Nearest Neighbor, Decision Tree, Support Vector Machine (SVM), Logistic Regression and Random Forest. The model with the best results will be optimized in order to fine tune it and compare the difference with the standard values.

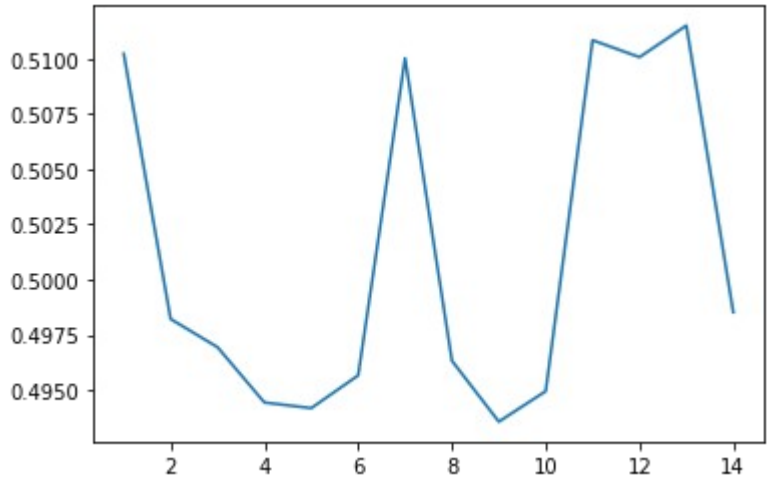
One of the things we have to do before training the model is to relabel our target variable as 0 for collisions with severity involving only property damage and 1 for those involving injuries. After that, it is important to select variables which can perform the best in the classification process and 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 universe of observations 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, we can iterate 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 we have relied in the accuracy score to select the best k.



Selecting 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. We 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

knn summary report

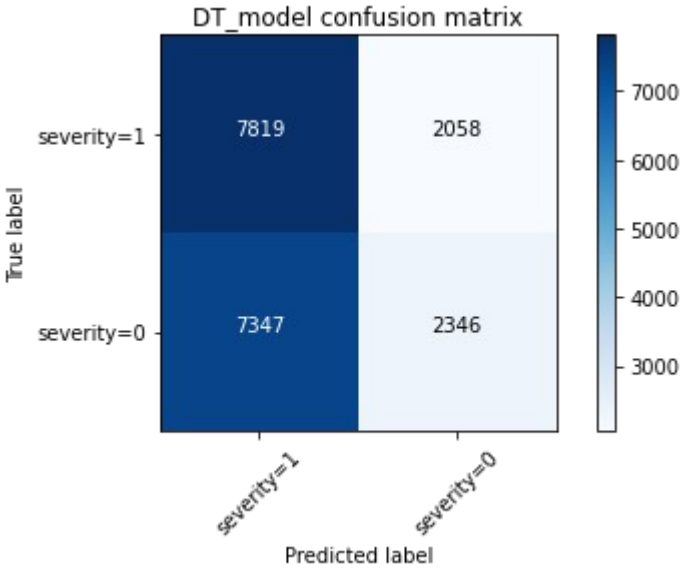
Decision Tree Modeling

Besides the default parameters, we have indicated to use the entropy criterion to create the many 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

Decision Tree summary report

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



Decision Tree confusion matrix

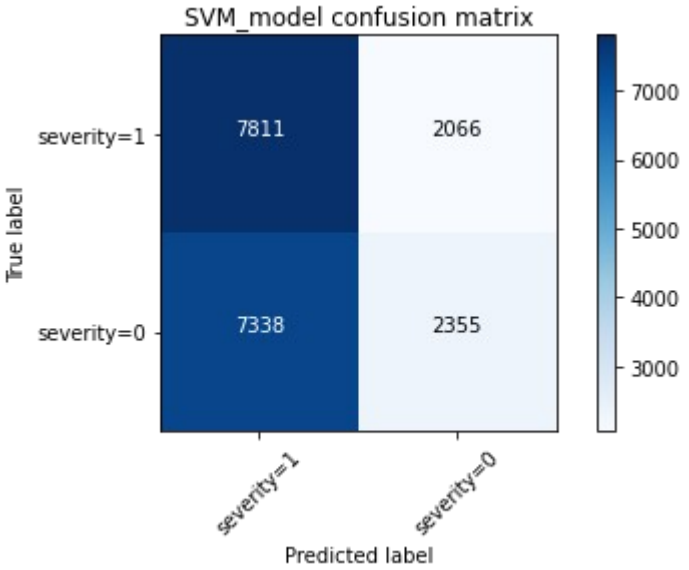
Support Vector Machine Modeling

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

SVM classification report

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.

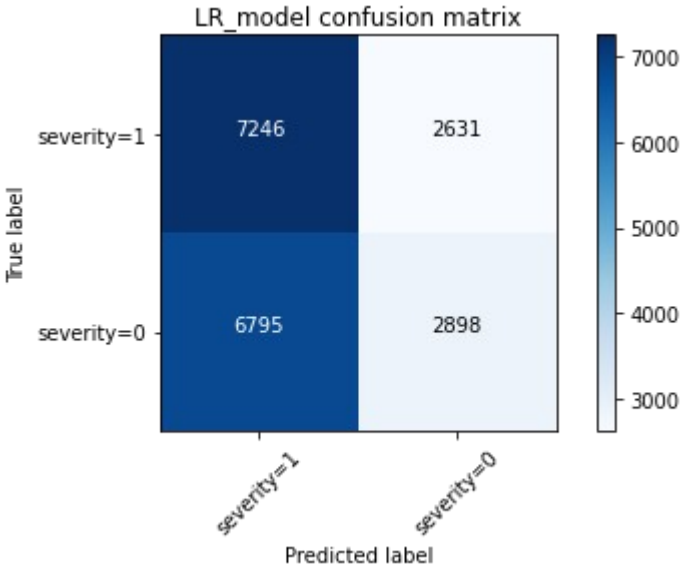


Logistic Regression Modeling

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 chance of over-fitting gets reduced. 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

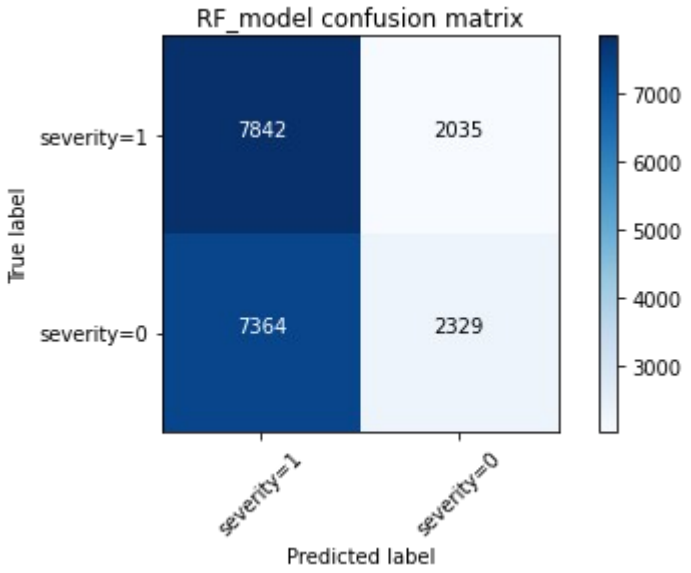
The Random Forest algorithm introduce randomness in the classifier construction when compared with traditional Decision Tree classifiers. The prediction of the ensemble is given as the averaged prediction of the individual classifiers.

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 same manner. 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

Summary

Based on the results obtained after the different the training of the analyzed models, the results are displayed in the table below. Briefly, the results show 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

Discussion

Conclusion