**Severity Prediction of An Accident Report**

**Introduction**

When driving, a bunch of factors such as weather conditions, road conditions, whether or not it is due to inattention and so on. In addition, the number of different type of vehicles involved in the collision is also a crucial indicator to assess the severity of this collision. If there is something in place that could warn people, given the weather and the road conditions about the possibility of people getting into a car accident and tell us how severe it would be, it is better for people to drive more carefully or even change the travel in advance.

Therefore, what we are going to do is to build an appropriate classification model so that we are able to predict the severity of a collision using these existing attributes. Many people will be interested in this project, for instance, drives themselves could get the information of how severe the accident would be, police could prepare in advance if the accident can’t be avoided in advance, the ambulances would be arranged given the prediction of the accident as well so that we can reduce the probability of casualties.

**Data Description**

There is a spreadsheet (.csv file) can be downloaded which contains both the labels and the prediction features. There are total 194673 records from 2004 to 2020.

The summary of this data set is provided by the GISWEB. SEVERITYCODE is a code that corresponds to the severity of the collision, 1 is prop damage and 2 represents there are some injuries. PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT and VEHCOUNT represent the total number of people, pedestrians, bicycles and vehicles involved in the collision respectively. JUNCTIONTYPE is the category of junction at which collision took place. Some attributes that are related to the drivers are as follows: INATTENTIONIND tells us whether or not the collision was due to inattention. UNDERINFL is whether or not the driver involved was under the influence of drugs or alcohol. SPEEDING denotes if speeding was a factor in the collision or not.

What’s more, there are some attributes concerning nature environment such as WEATHER, ROADCOND and LIGHTCOND.

**Methodology**

Data Exploration

* Eliminate useless features

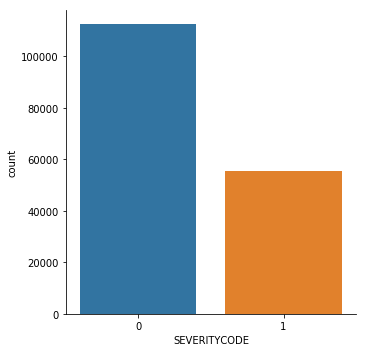
There are total 194673 observations and 37 features in the dataset. Features: X, Y, INCKEY, COLDETKEY, REPORTNO, STATUS, ADDRTYPE, INTKEY, LOCATION, EXCEPTRSNCODE, EXCEPTRSNDESC, SEVERITYCODE.1, SEVERITYDESC, INCDATE, INCDTTM, SDOT\_COLCODE, SDOT\_COLDESC, ST\_COLCODE, ST\_COLDESC, SDOTCOLNUM, SEGLANEKEY, CROSSWALKKEY are some general or descriptive features, so we will eliminate them in our prediction model.

* Missing Values

COLLISIONTYPE, JUNCTIONTYPE, WEATHER, ROADCOND, LIGHTCOND are 4 features that has NA in observations, so these records will be dropped. After dropping these rows, there are total 167980 observations without missing values.

* Imbalance checking

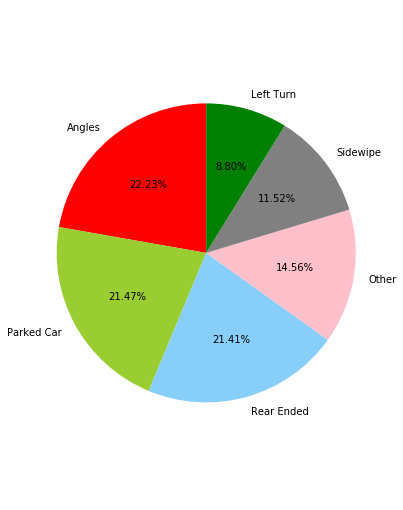
The label SEVERITYCODE has 112477 code1 and 55503 code2. Prop damage is almost twice as much as collisions with injury. Obviously, it is an imbalanced dataset.



* Feature Visualization

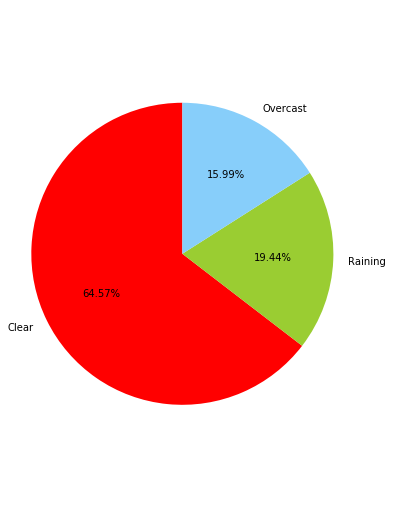
1. COLLISIONTYPE

There are 6 types of collision that result in over 10,000 observations, they are “Angles”, “Parked Car”, “Rear Ended”, “Other”, “Sideswipe” and “Left Turn”. Here is the pie chart that demonstrates the percentage of each type.



1. WEATHER

“Clear”, “Raining” and “Overcast” are the three top types of weather that contribute to the collision. The pie chart of weather is as followed:



* Encoding Method

In many Machine-learning or Data Science activities, the data set might contain text or categorical values (basically non-numerical values). Decision trees can handle categorical values very well but most of the algorithms expect numerical values. There are 2 main methods to convert categorical values into numerical values: One-Hot-Encoding and Label-Encoder.

1. Label Encoder

“INATTENTIONIND”, “UNDERINFL”, “PREDROWNOTGRNT”, “SPEEDING” and “HITPARKEDCAR” are columns of “category” datatype.

1. One-hot Encoding

“COLLISIONTYPE”, “JUNCTIONTYPE”, “WEATHER”, “ROADCOND” and “LIGHTCOND” are categorical features with more than two levels. One-hot Encoding is more appropriate to address this kind of features since label encoding has the disadvantage that numeric values can be misinterpreted by algorithms as having some sort of hierarchy or order in them. These features will be converted into new columns and assigned a 1 or 0 value to the column using One-hot Encoding.

So far, all the relevant features has been converted successfully and will be incorporated into the classification model to predict the severity of the collision.

**Machine Learning Methods**

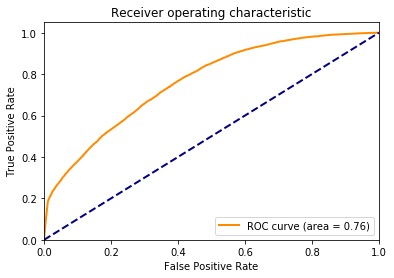
Since it is a imbalanced dataset, BalancedBaggingClassifier from *imblearn* library will be used to address this issue. It is a machine learning method that includes an additional step to balance the training set at fit time using a RandomUnderSampler. It allows the resampling of each subset of the dataset before training each estimator of the ensemble. Using this algorithm, a classifier can be trained without having to undersample or oversample manually before training.

Then it is important to tune the parameters in BalancedBaggingClassifier. Using Decision Tree Classifier as the base estimator model, parameters are tuned using GridSearch with 5-fold cross validation in the training set. The best number of base estimators in the ensemble (n\_estimators) is 500, the best number of samples to draw from X to train each base estimator(max\_smaples) is 0.5, the best number of features to draw from X to train each base estimator(max\_features) is 0.3.

**Results**

After applying the whole well- trained model on the test set, the True Negatives(Severity=1 is detected correctly) is 22420, False Negatives(Severity=1 is detected by mistake) is 4957, True Positives(Severity=2 is detected correctly) is 11547 and False Positives(Severity=2 is incorrectly detected) is 11470.

The evaluation method to assess the model is AUC value. AUC (Area Under The Curve) tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is. In addition, the ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis. Finally, the classification model in this project has a AUC=0.763 and the corresponding ROC is as followed:



**Discussion**

We could see that the AUC as well as the ratio of incorrect classification is not that satisfying, especially the True Positives and False Positives is almost the same. As a result, we may get better model selection when we balance the data manually. In terms of the classification model, other models may be applied such as XGBoost, BalancedRandomForest and so on.

What’s more, the model above is mostly rely on the count of previous accidents on the road segment as we can see from the feature importance of the accident count feature. This is not an issue for accident prediction, but it does not help to understand why these roads are particularly dangerous.

**Conclusion**

In this study, we conducted an analysis of the severity of collision in the city using open data provided by GIS. Using the train dataset, we built the severity of collision prediction model using balanced decision tree algorithm while tuning the best parameters. Our best model can predict 77% of the severity of collision. Moreover, we believe that our model can be reproduced easily for others cities under the condition that similar datasets are available. Our analysis is rather meaningful that it helps drivers to avoid severe collision as well as the relevant staff to make preparations in advance given the conditions of weather, road conditions and other attributes.