**Car Accident Severity Analysis (IBM Capstone Project)**

**Introduction**

Accidents in traffic lead to associated fatalities and economic losses every year worldwide and thus is an area of primary concern to society from loss prevention point of view. According to preliminary estimates from National Highway Traffic Safety Administration (NHTSA), 36,120 people died in motor vehicle crashes in [2019](https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812946), down 1.2 percent from 36,560 in 2018. Among all countries, the USA has the largest economic burden of road injuries of $487 billion, followed by China ($364 billion) and India ($101 billion); according to a research journal published by THE LANCET.

Reducing traffic accidents is an important public safety challenge around the world. Accident prediction is important for optimizing public transportation, enabling safer routes, and cost-effectively improving the transportation infrastructure, all in order to make the roads safer.

**Major Stakeholders**

* Travelers
* Insurance Companies
* State Health Department
* Emergency Services
* Infrastructural Development Authorities
* Families of the Travelers
* Taxpayers

**Problem**

There is a lack of awareness amongst travelers regarding the risks they might be facing while taking certain routes, crossing certain areas, driving at a specific speed, driving on a specific road, and being inattentive while driving, etc. High-accident-prone areas are seldom inspected with regards to road maintenance, and deployment of additional emergency services personnel, causing additional damage caused by road accidents.

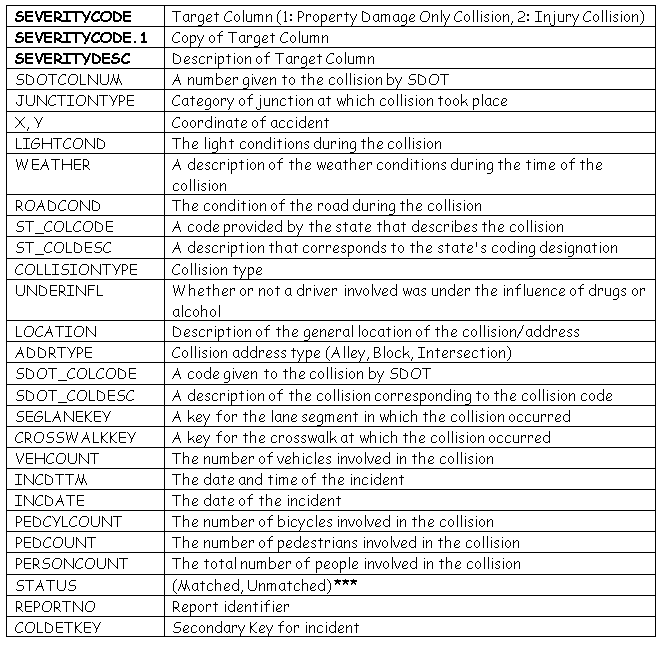
**Goal**

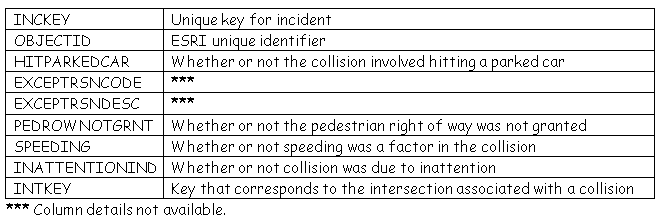
This project aims to predict whether an accident that happens under a specific set of circumstances will be an accident limited to property damage or if it will include some form of physical injury to the driver and/or the passengers. The goal of accident prediction is usually to provide a measure of the risk of accidents at different points in time and space. The occurrence of an accident is the label used to train the model, and the proposed model can be used to identify where and when the risk of accident is significantly higher than average in order to take actions to reduce that risk.

**Data**

All the collision data used in this analysis is taken from ArcGIS, which was provided by Seattle Police Department and recorded by traffic records. The data provided is that of collisions which took place in the city of Seattle, from year 2004 till present.

Mentioned below is list of features that was available in the raw data:





There are in total 38 data columns in the dataset including the 3 target related columns. We will keep various aspects in mind while deciding the importance of a particular column or the transformation it may need before we feed it to the model.

Some of the given data columns are features related to or identifying a single particular accident, thus may not be very much useful for our predictive analysis. These features include:

*SDOTCOLNUM, Coordinates, LOCATION, INCDTTM, INCDATE, REPORTNO, COLDETKEY, INCKEY, OBJECTID.*

There are some description columns for a given code. Columns *ST\_COLDESC, SDOT\_COLDESC and EXCEPTRSNDESC* are description columns for code which is already specified in the given dataset.

There are also data columns which has missing data in abundance. *Column EXCEPTRSNCODE,*EXCEPTRSNDESC,*PEDROWNOTGRNT, SPEEDING, INATTENTIONIND*and*INTKEY* have more than 50% of data missing. Although few of these columns can be very crucial indicator of collision severity, it would be misguiding to use it with so many missing rows and very difficult to fill in these categorical values.

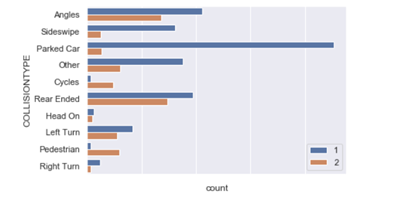
Columns mentioned in all the three categories above will not be used in the model that we are going to build. Most of the columns that remains are categorical and will require one-hot and label encoding before we can use them as a feature for our model.

**Methodology**

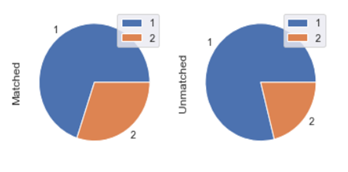
**Exploratory Data Analysis**

First part of the process will be to explore the data and understand that how a particular data column is distributed.

Most of our data columns are categorical and we need to know that how affect the severity of the accident.



Frequency of Property Damage Only Collision and Injury Collision with respect to collision type feature



Class distribution of ‘Matched’ and ‘Unmatched’ categories of status variable

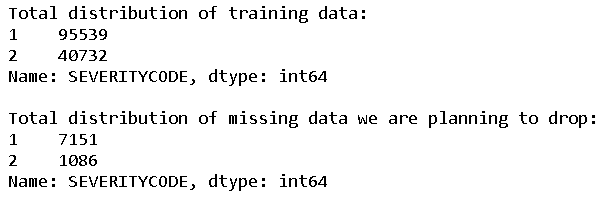
There are low cardinality categorical variables with 6–7 categories, moderate cardinality categorical variables with 40–70 categories and very high cardinality categorical variable with 1500+ categories.

**Feature Engineering**

Mostly all the variables (except the features which defines the number of people, vehicles etc.) are nominal features; i.e., features where the categories are only labelled without any order of precedence. Preferred encoding for these categories is One-Hot encoding. However, One-Hot encoding will generate around 1500 data columns for just one high cardinality categorical variable, which will be very expensive to work with.

We can get over this hurdle by using feature hashing. Feature hashing is an encoding technique which is used to encode high cardinality feature by hashing them. By this we can pull down the number of encoded data columns to 32–64 even for variables with >1500 categories.

Distribution of all missing data in the training set was found to be:

Distribution of all missing data in the training set

As the class proportion is not getting much affected by dropping these data rows, we will proceed to do so.

After the process of feature hashing and one-hot encoding, we obtain in total 208 feature columns. We are using Random Forest to get the feature importance, eliminating 40 least important features and correlation matrix to detect >90% correlations.

After removing the least important and highly correlated features we are left with 160 features to train the model with.

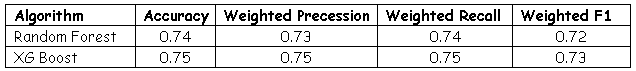
**Modelling**

As it was clear from above analysis that we have had a skewed dataset. This resulted in a low recall on class 2 and as a result low F1 score.

To solve this problem, we used smote to oversample the rare class and generated the cross-validation score again. While doing oversampling we have to keep in mind that oversampling should be done on each iteration of cross-validation and not on the whole training set.

As a result, we observed that although the recall on class 2 and F1 increased a little bit, it decreased the accuracy too. Considering the increase in computational expense due to increased data, oversampling didn’t prove to be worth the effort in this case.

We used the XG Boost Classifier to start with and plotted the learning curve to see if the model is overfitting the training data. We observed that converged training and validation error were close to each other, which means that we can use high variance algorithms like Random Forest, XG Boost and Support Vector Machine, and we can also use the high number of features that we are using.

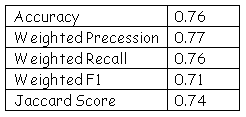
Cross-validation results for both the algorithms

As expected, we got the best performance from XG Boost Classifier. We will further try hyperparameter tuning to improve the performance.

**Results**

For final prediction we have to preprocess the whole test dataset. While encoding the feature columns we made sure that the one-hot encodings are same as the train set and feature hasher transformer used should be fitted on train data.

Following are the Final result on the test data:



Final Evaluation on Test data

**Discussion**

Many more analysis and methodologies can be added to this project as a future work. We haven’t used the coordinates. Those coordinates could result in some unforeseen clusters which could exponentially improve the study.

Further other encoding techniques can be used in place of feature hashing or feature hashing with different feature count can be used. The performance of these changes can be evaluated using cross-validation.

**Conclusion**

The results are satisfactory, but expectations were much higher. A lot of improvement can be done on class 2 predictions. Overall a lot of improvement can be observed from the basic model.