Predicting Car Accident Severity

Introduction

Road traffic accidents are common safety problem around the world. This automotive accidents result in over result in over 30,000 fatalities in the United States annually. Road traffic accidents are estimated to cost the US economy approximately $810bn per. Identifying the factors which influence accident severity is therefore of paramount importance.

In an effort to reduce the frequency of car collisions in our community, this project will leverage existing accident data to predict the different accidents' severity given the current weather, road and visibility conditions. When conditions are bad, this model will alert drivers to remind them to be more careful.I will use different supervised machine learning algorithms and select the machine learning model that gives the highest prediction accuracy.

The study of building this kind of model will be of significance to a lot of stakeholders and beneficiaries: (1) town/city planners, who may be able to use the model to inform their road planning and traffic calming strategies; (2) emergency service responders, who may be able to use the model to predict the severity of an accident based on information that's provided at the time the accident is reported in order to optimally allocate resources across the city, and (3) traffic police officers.

Data

A comprehensive dataset of over 190,000 observations collected occurring between 2004–2019 in the Seattle city area was obtained from the Seattle Open Data Portal. The dataset has almost 40 columns describing the details of each accident including the weather conditions, collision type, date/time of accident and location. To accurately build a model to prevent future accidents and/or reduce their severity, we will use the following attributes — ADDRTYPE, WEATHER, ROADCOND, VEHCOUNT, PERSONCOUNT.

Methodology

The following will be covered:

- exploratory data analysis
- inferential statistical testing
- machine learnings

Data Cleaning

Firstly, the selected data *INCDTTM* and *INCDATE* column values were converted to the appropriate datetime format, for better manipulation. This generated a number of missing values, in particular from the *INCDTTM* where there are a lot of missing time values. We decided to keep these entries during the EDA, but eventually we remove these missing value entries when we eventually decided that the time feature was of importance to our analysis, and that the SEVERITYCODE distribution within the entries with missing time values were similar to the overall dataset true distribution.

To resolve the other missing values coming from *ROADCOND, LIGHTCOND, WEATHER*, we used the filtered out those that had 'Unmatched' values in the *STATUS* column as these had a high proportion of missing *ROADCOND, LIGHTCOND, WEATHER* values. For the remaining entries with null values, they constitute a small proportion <1% of the total dataset and were also removed.

For the remaining missing values in *ADDRTYPE, JUNCTIONTYPE*, we found that a proportion of them had missing values in both *ADDRTYPE, JUNCTIONTYPE* and these were dropped as they will not provide much value. We would that a large proportion of the remaining missing values in the *JUNCTIONTYPE* column contain the *ADDRTYPE* 'Block' value. As it is unknown what the led to the missing values, we decided to relabelled all remaining missing values in these two columns as 'Unknown'.

Data Processing

As we've kept a couple of categorical features, we utilised both Label Encoding during the EDA to visualise data correlation, and One-Hot Encoding on the final feature set so that they can be input into the model. To enable ease of manipulation, the feature names were all also converted to lowercase.

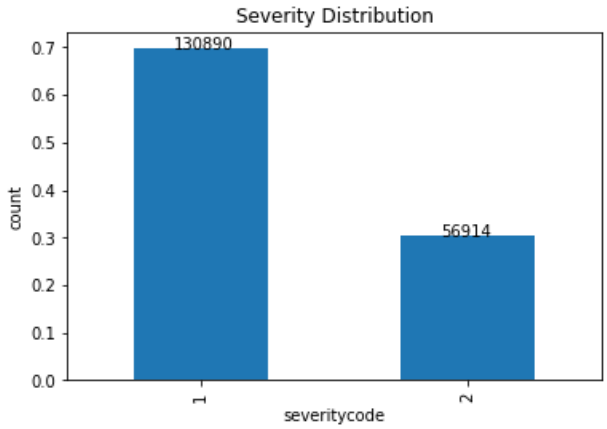
We created a *day*, *month* and *hour* feature from the *incdate* and *incdttm* features for EDA and model buildings. Other features such as *severitycode*, *hitparkedcar* were all relabelled to appropriate values of 0 and 1 according.

Exploratory Data Analysis

Target Variable: Severity

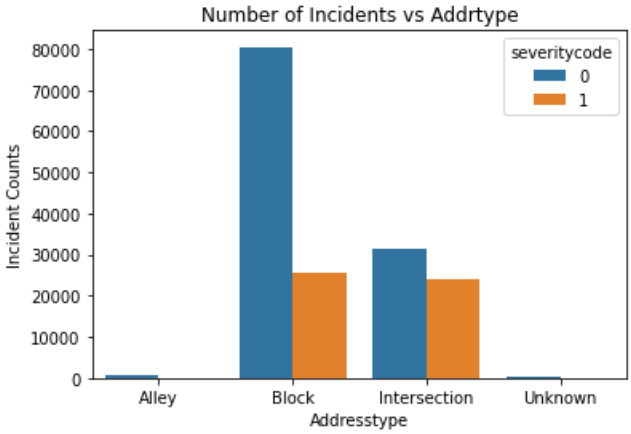
The main target variable in this case is accident severity, represented by the feature column *severitycode*. It was found that the dataset is imbalanced, with the proportion of the minority class corresponding to *severitycode* of 1 at about 30%.

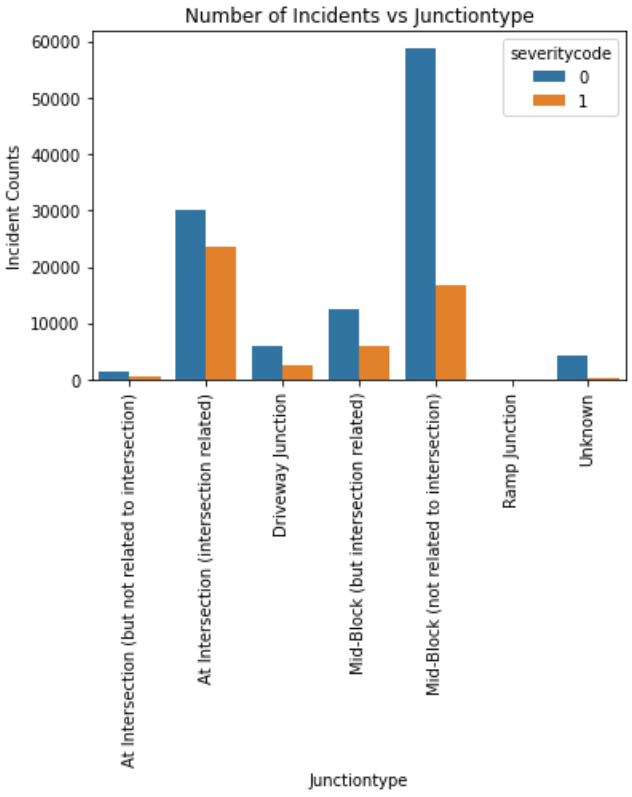
While this may be the naturally occurring distribution of incident severity, down sampling of the majority class or up sampling of the minority class can be considered as strategies to improve the performance of the model. However, as a start, we will build the model with the original dataset with the true distribution.



Location Conditions

We see that there are mild correlations between *severitycode* and the *junctiontype* and *addrtype* features (Figure 7). Overall, we see a huge proportion of incidents occurring at the blocks, in particular, mid-block, unrelated to the intersection (Figure 6). However, we also see that there is a strong correlation between *junctiontype* and *addrtype* in Figure 7. We thus dropped one feature (*addrtype* in this case) to reduce multicollinearity and skewed results, given that the information provided by *addrtype* is also captured by the *junctiontype*.





Conclusion

Most crashes happened in clear, dry, and bright conditions. Most days are clear, dry, and bright, so it's no surprise that most car crashes occur under these conditions. I also found out that crashes with a distracted driver or an impaired driver are statistically more likely to result in injury, which is also not a surprise. The results of the data indicate to city officials that they should ask drivers to be more alert in ideal conditions

We see an improvement in the Jaccard and Recall scores of all models after rebalancing the training dataset. With a balanced dataset, the GBC model has improved and performed better in all aspects than the RF model, with the LR model close in predictive performance. A closer look into the classification reports above show that most model suffer from poor precision for the severe case, i.e. when *severitycode* is 1, i.e. there is a tendency for the model to make False Positives.

Overall, we see that GBC returns the best performance compared to LR and RF, boasting a recall of 0.70 vs 0.64 and 0.63 from LR and RF respectively. The poor Jaccard score for all models however meant that the models tend to fail in predicting a large portion of the dataset correctly, likely due to poor precision.