**Predicting the Severity of Accidents in the City of Seattle**

David Arthur

September 28, 2020

1. **Introduction**
   1. **Background**

According to the [WHO](https://www.who.int/health-topics/road-safety#tab=tab_1) reports Road traffic injuries are now the leading killer of people aged 5-29 years. In addition to the human suffering caused by road traffic injuries, there is also heavy economic burden on victims and their families, both through treatment costs for the injured and through loss of productivity of those killed or disabled. More broadly, road traffic injuries have a serious impact on national economies, costing countries 3% of their annual gross domestic product.

The WHO also states a number of factors that increase both the risk of road traffic crashes and the risk of death or injury they result in such as; driving speed which speed significantly increases both the likelihood of a crash occurring, and the severity of its consequences.

Driving under the influence of alcohol or other psychoactive substances presents significant risk factor for road traffic injuries

Other significant risk factors are the non-use of motorcycle helmets, seatbelts, and child restraints, distraction, unsafe vehicles and unsafe road infrastructure can negatively impact safety on the roads, inadequate post-crash care, inadequate law enforcement of traffic laws

* 1. **Problem**

The objective of this project is to explore the data, find features important in predicting the severity of accidents, build models for predicting the severity of an accident and finally selecting the top performing model using various classification metrics with the hope that the insights gathered from this project can help the City of Seattle's Department of Transportation planning and policy making to reduce the number of accidents especially severe accidents.

1. **Data Acquisition and cleaning**

**2.1 Data sources**

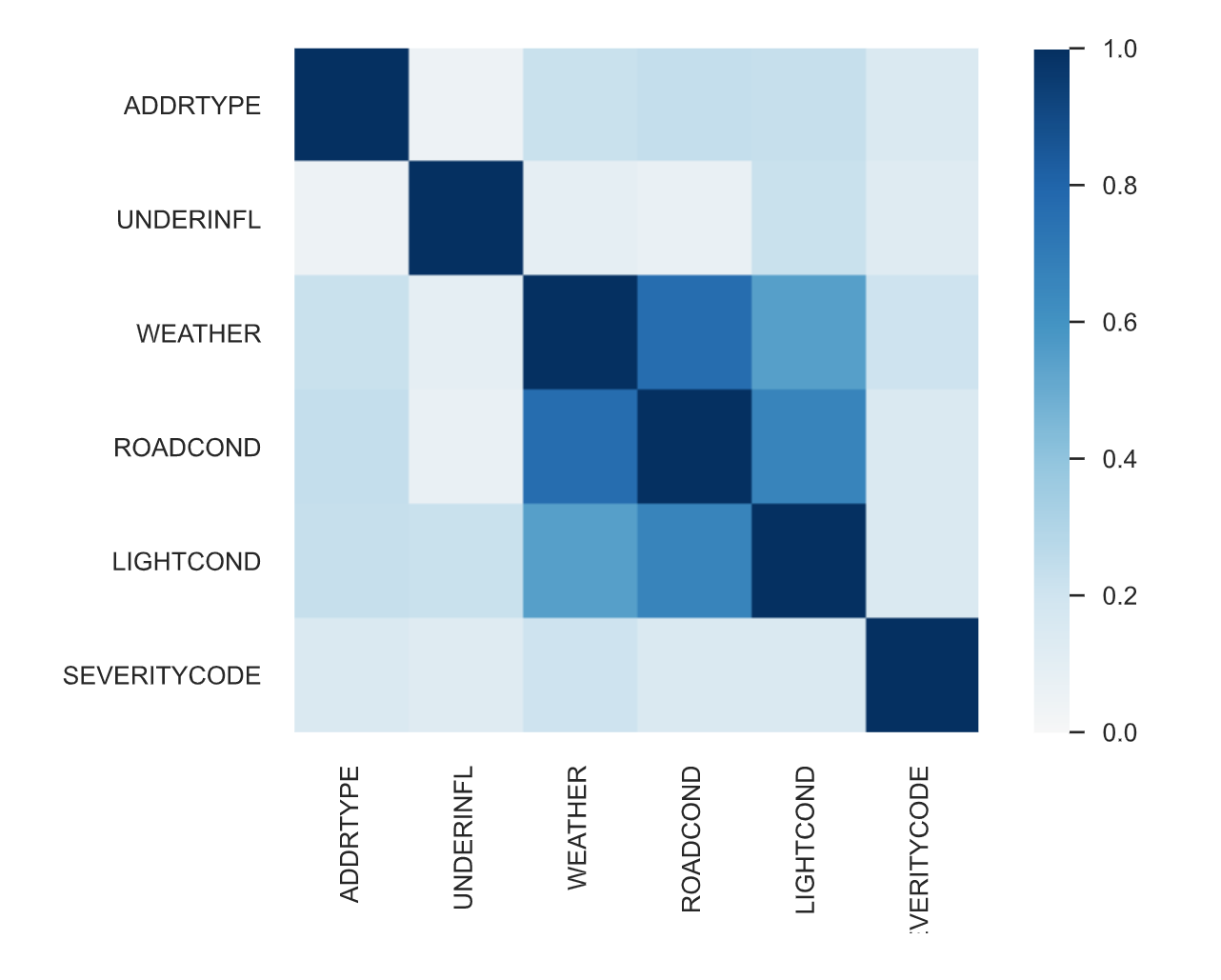
The Data for this project was obtained from the City of Seattle's Open Data Portal. The Portal makes data generated by the City openly available to the public for the purpose of increasing the quality of life for the residents, increasing transparency, accountability and comparability, promoting economic development and research, and improving internal performance management. The data is updated weekly and can be found at the Seattle Open GeoData Portal.

**2.2 Data cleaning and feature selection**

From initial descriptive statistics it is shown the data set has 221389 observations and 40 features including the dependent variable SEVERITYCODE. Preliminary cleanup exercise would be to remove observations with missing values for SEVERITYCODE. Examining the values of SEVERITYCODE we have 5 unique values including one representing unknown, since it is a goal to predict a known level of SEVERITY, observations with unknown severity are removed.

Features ending with DESC which are meant to provide text description for the variables ending with CODE are also removed. Features ending with COUNT representing the counts of persons, pedestrians, bicycles, vehicles since they provide more information on the after effect of a collision not the possible factors for a collision. Features ending with KEY are unique ID columns and are removed. Highly correlated features are also removed since they may provide redundant information and may not improve predictive performance.

The subset of features selected for building the model was ADDRTYPE, INCDTTM, JUNCTIONTYPE, INATTENTIONIND, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, SPEEDING, SEVERITYCODE.

Correlation plot below gives us an idea of how the features are related to each other as well as the dependent variable SEVERITYCODE

1. **Predictive Modeling**

Three models were built to predict SEVERITYCODE with the selected feature set. These models are the Decision Tree Classifier, the Random Forest and the Logistic Regression models. The implementation of the models exposes the difficulty models encounter when predicting severely imbalanced dependent variables. This problem is catered to by balancing the classes by either oversampling minority classes, under-sampling majority classes, applying SMOTE or using models that are not sensitive to imbalanced classes.

**3.1 Decision Tree Classifier**

The Decision tree classifier produced the performance results shown below. It can be seen the model could not predict examples in the minority classes and achieves an overall accuracy of 66% by just predicting the two majority classes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| class | precision | recall | f1-score | support |
|  |  |  |  |  |
| 1 | 0.66 | 0.99 | 0.79 | 37934 |
| 2 | 0.46 | 0.02 | 0.04 | 18551 |
| 2b | 0 | 0 | 0 | 983 |
| 3 | 0 | 0 | 0 | 110 |
|  |  |  |  |  |
| accuracy |  |  | 0.66 | 57578 |
| macro avg | 0.28 | 0.25 | 0.21 | 57578 |
| weighted avg | 0.58 | 0.66 | 0.54 | 57578 |

**3.2 Decision Tree Classifier With Balanced Classes**

This Decision Tree Classifier model has the class weight parameter set to balance to accommodate for the severe imbalances. This model achieves an overall accuracy of 55% mainly because it fails to classify a lot of the minority classes, overall accuracy drops because of this

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| class | precision | recall | f1-score | support |
|  |  |  |  |  |
| 1 | 0.73 | 0.62 | 0.67 | 37934 |
| 2 | 0.42 | 0.45 | 0.44 | 18551 |
| 2b | 0.02 | 0.02 | 0.02 | 983 |
| 3 | 0.01 | 0.39 | 0.02 | 110 |
|  |  |  |  |  |
| accuracy |  |  | 0.55 | 57578 |
| macro avg | 0.3 | 0.37 | 0.29 | 57578 |
| weighted avg | 0.62 | 0.55 | 0.58 | 57578 |

**3.3 Random Forest Classifier**

The Random Forest classifier produced the performance results shown below. It can be seen the model could only predict examples in the majority classes and achieves an overall accuracy of 66% by just predicting all examples to be in the majority classes which happens to be 66% of the overall data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| class | precision | recall | f1-score | support |
|  |  |  |  |  |
| 1 | 0.66 | 1 | 0.79 | 37934 |
| 2 | 0 | 0 | 0 | 18551 |
| 2b | 0 | 0 | 0 | 983 |
| 3 | 0 | 0 | 0 | 110 |
|  |  |  |  |  |
| accuracy |  |  | 0.66 | 57578 |
| macro avg | 0.16 | 0.25 | 0.2 | 57578 |
| weighted avg | 0.43 | 0.66 | 0.52 | 57578 |

**3.4 Random Forest Classifier With Balanced Classes**

This Random Forest Classifier model has the class weight parameter set to balance to accommodate for the severe imbalances. This model achieves an overall accuracy of 47% mainly because it fails to classify a lot of the minority classes, overall accuracy drops because of this

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| class | precision | recall | f1-score | support |
|  |  |  |  |  |
| 1 | 0.73 | 0.63 | 0.68 | 37934 |
| 2 | 0.42 | 0.14 | 0.21 | 18551 |
| 2b | 0.02 | 0.35 | 0.05 | 983 |
| 3 | 0.01 | 0.35 | 0.02 | 110 |
|  |  |  |  |  |
| accuracy |  |  | 0.47 | 57578 |
| macro avg | 0.3 | 0.37 | 0.24 | 57578 |
| weighted avg | 0.62 | 0.47 | 0.52 | 57578 |

**3.5 Logistic Regression Classifier**

The Logistic Regression classifier produced the performance results shown below. It can be seen the model could not predict examples in the minority classes and achieves an overall accuracy of 66% by just predicting the two majority classes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| class | precision | recall | f1-score | support |
|  |  |  |  |  |
| 1 | 0.66 | 0.98 | 0.79 | 37934 |
| 2 | 0.44 | 0.04 | 0.07 | 18551 |
| 2b | 0 | 0 | 0 | 983 |
| 3 | 0 | 0 | 0 | 110 |
|  |  |  |  |  |
| accuracy |  |  | 0.66 | 57578 |
| macro avg | 0.28 | 0.25 | 0.22 | 57578 |
| weighted avg | 0.58 | 0.66 | 0.54 | 57578 |

**3.6 Logistic Regression Classifier With Balanced Classes**

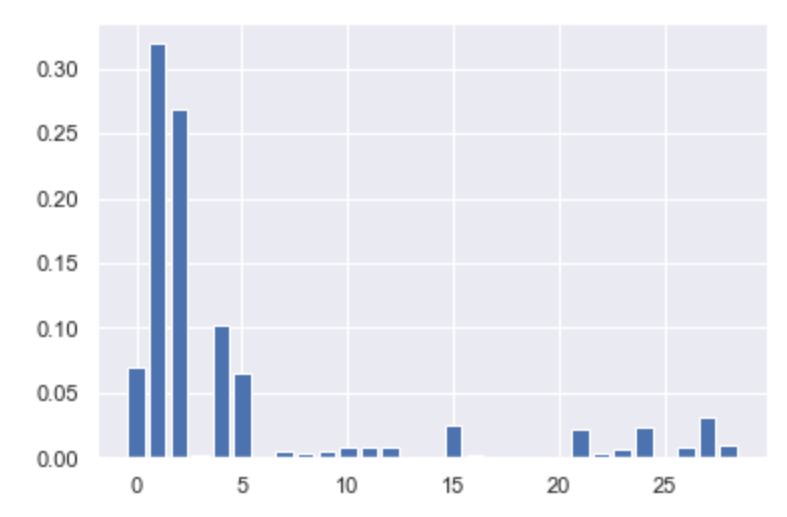
This Logistic Regression Classifier model has the class weight parameter set to balance to accommodate for the severe imbalances. This model achieves an overall accuracy of 46% mainly because it fails to classify a lot of the minority classes, overall accuracy drops because of this

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| class | precision | recall | f1-score | support |
|  |  |  |  |  |
| 1 | 0.73 | 0.62 | 0.67 | 37934 |
| 2 | 0.43 | 0.15 | 0.22 | 18551 |
| 2b | 0.02 | 0.32 | 0.04 | 983 |
| 3 | 0.01 | 0.41 | 0.02 | 110 |
|  |  |  |  |  |
| accuracy |  |  | 0.46 | 57578 |
| macro avg | 0.3 | 0.38 | 0.24 | 57578 |
| weighted avg | 0.62 | 0.46 | 0.52 | 57578 |

**3.7 Feature Importance**

The feature importance scores and plot below gives an idea of how important a feature is in contributing to predicting the independent variable. This data is generated from the Random Forest Model

|  |  |  |
| --- | --- | --- |
| Feature: 0 | INATTENTIONIND | Score: 0.06953 |
| Feature: 1 | UNDERINFL | Score: 0.31929 |
| Feature: 2 | SPEEDING | Score: 0.26851 |
| Feature: 3 | Alley | Score: 0.00263 |
| Feature: 4 | Block | Score: 0.10213 |
| Feature: 5 | Intersection | Score: 0.06496 |
| Feature: 6 | Blowing Sand/Dirt | Score: 0.00005 |
| Feature: 7 | Clear | Score: 0.00547 |
| Feature: 8 | Fog/Smog/Smoke | Score: 0.00407 |
| Feature: 9 | Overcast | Score: 0.00469 |
| Feature: 10 | Partly Cloudy | Score: 0.00867 |
| Feature: 11 | Raining | Score: 0.00757 |
| Feature: 12 | Severe Crosswind | Score: 0.00826 |
| Feature: 13 | Sleet/Hail/Freezing Rain | Score: 0.00062 |
| Feature: 14 | Snowing | Score: 0.00045 |
| Feature: 15 | Dry | Score: 0.02579 |
| Feature: 16 | Ice | Score: 0.00146 |
| Feature: 17 | Oil | Score: 0.00001 |
| Feature: 18 | Sand/Mud/Dirt | Score: 0.00005 |
| Feature: 19 | Snow/Slush | Score: 0.00049 |
| Feature: 20 | Standing Water | Score: 0.00016 |
| Feature: 21 | Wet | Score: 0.02228 |
| Feature: 22 | Dark - No Street Lights | Score: 0.00416 |
| Feature: 23 | Dark - Street Lights Off | Score: 0.00603 |
| Feature: 24 | Dark - Street Lights On | Score: 0.02415 |
| Feature: 25 | Dark - Unknown Lighting | Score: 0.00000 |
| Feature: 26 | Dawn | Score: 0.00757 |
| Feature: 27 | Daylight | Score: 0.03156 |
| Feature: 28 | Dusk | Score: 0.00939 |



From the scores and plot we can see UNDERINFL, SPEEDING, Block, Intersection, Dry, Wet, Dark - Street Lights On, Daylight were considered by the model to be important in predicting the SEVERITYCODE.

**4.0 Conclusion**

In this study various features were studied to investigate their influence in predicting the Severeness of a collision (SEVERITYCODE). Whether or not the individuals were under influence , speeding, whether or not the address type is a block or intersection, whether or not the road conditions were wet or dry and whether or not it was dark with street lights on or off or it was day were found to be important in predicting the severity of a collision. The features listed can be considered in policing making by the Seattle Department of transportations and measures taken to ensure that conditions are ideal to minimize the probabilities of severe collisions

**5.0 Future Directions**

The models considered in this study could not better the benchmark accuracy which is the percentage of the majority class. Areas that can be considered to improve this is using imputation methods like using mean or median to replace missing values which were predominant in the dataset or the MICE algorithm. Also other predictive models could be explored and various hyperparameters chosen to try and improve predictive performance.