**IBM Data Science Professional Certificate**

**Capstone Project – Predict Car Accident Severity**

**in Seattle City**

1. **Introduction**

Driving is an inseparable part of our daily life - It brings convenience to us, but at the same time we are always exposed to the risk of potential car accidents every time when we drive on the road. To protect our self from running into accidents which cause severe damages and the loss of life (the worst), we should be better off if we can know in advance about the possibility and severity of a car accident - which might occur given the weather, road condition and other factors. Predicting car accidents is important for all of us. Reducing the property and personal damages & loss is also beneficial for a safe and healthy society. For professional drivers who are exposed to a much higher risk of car accidents, they would be even more interested in predicting a car accident. Lastly, if we are visiting a city or place which we are not familiar with, if we can predict the severity of a car accident, then we will drive more safely to avoid the big accidents as much as possible.

1. **Data Source and Cleaning**
   1. **Data Sources**

This dataset is about car collisions occurred in Seattle City from year 2004 to present. It is comprised of 194,673 records with 38 attributes which some of them will be used to train and test the machine learning model for this project. This dataset has unbalanced labels because some of them have missing data, and it has both categorical and numerical types of data. The objective of this project is to build a classification model to predict the severity of a unknow (or new) car collision based on a set of features selected.

* 1. **Data Cleaning**

To balance the dataset for less biases and better predictability, several steps are taken for data cleaning.

2.2.1 For unbalance labels, remove rows without useful values. This step results in downsizing the dataset to 168,947 records from original size of 194,673.

1. Remove rows with missing values, including:

|  |  |
| --- | --- |
| **Columns** | **Records Removed** |
| ADDRTYPE | 1,926 |
| WEATHER | 5081 |
| ROADCOND | 5,012 |
| LIGHTCOND | 5,170 |

1. Remove rows with small values which are not predictive for modeling.

|  |  |
| --- | --- |
| **Columns** | **Records Removed** |
| VEHCOUNT | Values >6 |
| PERSONCOUNT | Values >10 |

3. Remove rows in columns 'WEATHER', 'ROADCOND', and 'LIGHTCOND'. which have values as 'Other', 'Unknown', and ' '.

|  |  |
| --- | --- |
| **Columns** | **Records Removed** |
| WEATHER | Unknown', 'Other', ' ' |
| ROADCOND | Unknown', 'Other', ' ' |
| LIGHTCOND | Unknown', 'Other', ' ' |

2.2.2 Convert categorial values to numerical values for columns including: 'WEATHER', 'ROADCOND', 'LIGHTCOND', and 'ADDRTYPE'.

* 1. **Feature Selection**

Since this dataset is about severity, the attribute 'SEVERITYDESC' is labelled as the target variable - y. Among the remaining 37 attributes, 6 of them are selected to the feature sets, including: 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'ADDRTYPE', 'PERSONCOUNT', and 'VEHCOUNT'. Other 31 attributes instead are dropped from the data frame file after cleaning.

1. **Methodology**
   1. **Exploratory Data Analysis**

In the dataset, the target variable 'SEVERITYDESC' has only two types of value - 'Property Damage Only Collision' vs. 'Injury Collision', with records of 136,485 vs.58,188 for each class. After data cleaning, the number of each class is reduced to 113,546 and 55,401, by 16.8% and 4.8% respectively. This indicates the data cleaning has unbalanced impact on the target variables, which results in records related to type I 'Property Damage Only Collision' being dropped much more than those related to type II 'Injury Collision'. The reason seems to be the higher severity of a collision, the more accurate and complete data are available in this dataset. This unbalance of data between different levels of severity of collisions would create some biases for this modeling, which as a result affect the predictability of it.

* 1. **Machine Learning Modeling**

This dataset is used to predict the possible outcome of a car collision. By applying classification algorithms to this dataset, a machine learning model is built to classify and predict the result. As this dataset contains mixed types of data - categorical and numerical values, KNN method is not an applicable approach. The Decision Trees method is used for modeling. The selected feature sets of this dataset are weather, road condition, light condition, address type, # of persons, and # of vehicles, and the target is the severity of collision which those features lead to. After splitting the dataset into training and testing part, the training part is used to build a decision tree which will predict the severity of a car collision.

1. **Results**

The Decision Trees classification approach is predictive for this project - with an accuracy rate of 72.2%. (max depth=4). Although the best depth is 7, after applying it to the max depth, the accurate rate is not increased much, while more depths for this case just overwhelm the decision tree graph, which complicates the decision tree analysis. Therefore, max depth is set to 4. There are two features ‘WEATHER’ and ‘ROADCOND’ is not applied in the Decision Tree Graph, the possible reason is the model chose the best 4 features to fit a tree of such levels. The severity of a car collision is highly related to 3 key features: ‘VEHCOUNT’ (number of vehicles), ‘PERSONCOUNT’ (number of persons), ‘ADDRTYPE’ (the address type).

|  |  |  |  |
| --- | --- | --- | --- |
| **Features** | **Descriptions** | **Severity** | **Entropy** |
| ADDRTYPE is not <=1.5 and VEHCOUNT is not <=2.5 | Car collision at an 'Intersection', 3 or more cars involved | Injury Collision | >0.90 |
| ADDRTYPE is <=1.5 | Car collision at 'Block' or 'Alley' | Property Damage Only Collision, | >0.90 |
| *\*\*\*Exception: predicted outcome with much high accuracy* | *two or less persons and two cars involved,* | *Property Damage Only Collision,* | *0.647* |
| VEHCOUNT is <=1.5 and PERSONCOUNT is not <=1.5 | Car collision involves only one or none of car, but two or more persons | Injury Collision | 0.531~0.783 |
| *\*\*\*Exception: predicted outcome with much low accuracy* | *light condition is dark* | *Injury Collision* | *0.965* |
| PERSONCOUNT is <=1.5 and not <=0.5 , and VEHCOUNT is<=0.5 | Car collision involves only one person, no car | Injury Collision | 0.097 |
| PERSONCOUNT is <=1.5 and not <=0.5 , and VEHCOUNT is <=1.5 and not <=0.5 | Car collision involves only one person, one car | Property Damage Only Collision, | 0.824 |
| PERSONCOUNT is <=0.5 and ADDRTYPE is not <=1.5 | Car collision involves nobody at 'Intersection' | Injury Collision | 0.816 |

1. **Discussion of Observation and Recommendations**

The Decision Tree classification model is trained based on a set of features from the dataset. How closely the target variable (severity) changes with those features shows if they are highly predictive variables or not. The Decision Tree model chose the number of vehicles ‘VEHCOUNT’ as the first attribute to split the dataset, but the ‘entropy’ is so high - 0.99, which results in impure nodes.

To increase the purity of the split data, more predictive feature with high confidence should be tested. after explore the dataset again, feature ‘SPEEDING’ was added to the feature set, and this time the model chose the number of person ‘PERSONCOUNT’ as the first attribute to split the dataset, but the accuracy rate is too low - 62%, which means the added feature has brought more biases to the model. The reason is because the feature ‘SPEEDING’ has too many missing values, with only 9,060 records with input value (compared to 168,947 records in the dataset (after data cleaning)).

Further dig into the relationship between the feature ‘SPEEDING’ and the target variable, you will see that 5,619 cases out of 9,060 records labeled ‘Y’ are ‘Property Damage Only Collision’, which means 62% of speeding cases results in less severe damage. However, the fact might be the opposite, because usually speeding is considered as a main cause of car accidents. Such an important feature ‘SPEEDING’ was dropped from the feature sets because of its incomplete data, which undermines the predictability of this model.

In additions, this dataset does not count all the pedestrians and cycles as the number of persons involved in a car collision. This dataset classifies severity into two types: Property Damage Only Collision and Injury Collison – which causes the lack of details about the severity of the injury. To build a more predictive machine learning model to predict the severity of a car collision, we need more accurate and better quality of data to be added to this dataset.

1. **Conclusion**

The performance of this Decision Tree ML modeling to predict the severity of a car collision is good. It provides insights about the main causes of severe car accidents:

1. The intersection is a dangerous place when injury collision usually happens, especially when 3 or more cars are involved.
2. When two cars collide not in an intersection, mostly likely the collision would be not severe, especially if just two or less persons involved (two drivers only or one of them).
3. Most of the severe collisions happens when less car (one or less) but more persons (two or more) are involved, and at the good light conditions (e.g., daylight). This finding turns upside down our traditional thinking about the impact of darkness on the occurrence of car collision. Based on this model, driving with good light (daylight, dawn, dusk) is prone to severe collision once it happens, especially at dusk it is the most dangerous time for safe driving. Maybe because usually people would drive more carefully in darkness.
4. When a car collision involves no vehicle but one person, 100% it would be an injury collision, which reminds us driving with extra caution at block and intersection.