**PERDICTING FOR ACCIDENT SEVERITY**

October 02, 2020

1. **Introduction**

**1.1 Background**

Facts are that cities and major highways around the globe have recorded high rate of road fatality resulting from accident severity. These accidents are majorly caused by human errors; violation of traffic rules; driving under the influence of alcohol; road and weather conditions and so on. Accident cannot be eliminated from our roads as we commute from one place to another, but the Severity can be mitigated if all the causes are analysed and solutions are proffered. This whole idea constitutes a brief description of this discuss.

* 1. **Problem**

The data provide an overview of the various incidents that have occurred, and factors that have influenced these incidents. These information include: accident severity whether it was low or high, location of the incident, severity impact, that is, whether it was an injury collision or property damage only collision, collision type, numbers of individuals involved, number of vehicles involved, road and weather conditions at the time of each incidents etc. This project is aimed at using data science to understand the impact of each of the influences mentioned above and how to help improve on this, in order to advert the critical loss of lives and properties upon accident severities.

**1.3 Data Description**

The data contain a record of 36 attributes which ranges from the location of the impact collision, factor before, during and after impact collision. These factors before impact collision include the road, weather, and light condition before the impact. Factors during impact collision include collision type, accident severity description, persons involved, number of vehicle involved, if the driver was under so sort of influence, and factors after collision include the level accident severity whether high or low and impact of collision on its surrounding.

* 1. **Interest**

The reduction in severity of accidents can be beneficial to the Public mostly which works towards improving those road factors and the car drivers themselves who may take precaution to reduce the severity of accidents.

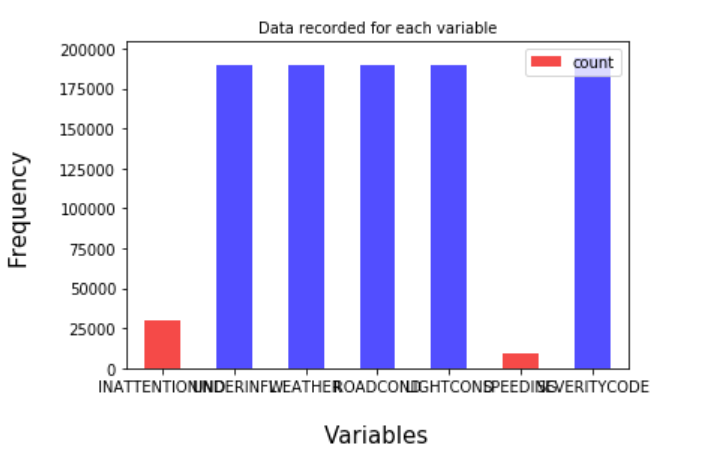
1. **Data Preprocessing**

**2.1 Data Cleaning**

There are a lot of problems with the data set keeping in mind that this is a machine learning project which uses classification to predict a categorical variable. The dataset has total observations of 194673 with variation in number of observations for every feature. First of all, the total dataset was high variation in the lengths of almost every column of the dataset. The dataset had a lot of empty columns which could have been beneficial had the data been present there.

These columns included pedestrian granted way or not, segment lane key, cross walk key and hit parked car. The models aim was to predict the severity of an accident, considering that, the variable of Severity Code was in the form of 1 (Property Damage Only) and 2 (Injury Collision) which were encoded to the form of 0 (Property Damage Only) and 1 (Injury Collision).

Furthermore, the Y was given value of 1 whereas N and no value was given 0 for the variables Inattention, Speeding and Under the influence. For lighting condition, Light was given 0 along with Medium as 1 and Dark as 2. For Road Condition, Dry was assigned 0, Mushy was assigned 1 and Wet was given 2. As for Weather Condition, 0 is Clear, Overcast is 1, Windy is 2 and Rain and Snow was given 3. 0 was assigned to the element of each variable which can be the least probable cause of severe accident whereas a high number represented adverse condition which can lead to a higher accident severity. Whereas, there were unique values for every variable which were either ‘Other’ or ‘Unknown’, deleting those rows entirely would have led to a lot of loss of data which is not preferred.

****

In order to deal with the issue of columns having a variation in frequency, arrays were made for each column which were encoded according to the original column and had equal proportion of elements as the original column. Then the arrays were imposed on the original columns in the positions which had ‘Other’ and ‘Unknown’ in them. This entire process of cleaning data led to a loss of almost 5000 rows which had redundant data, whereas other rows with unknown values were filled earlier.

**2.2 Feature Selection**

A total of six (6) features were trained while the target / label variable being Severity code.

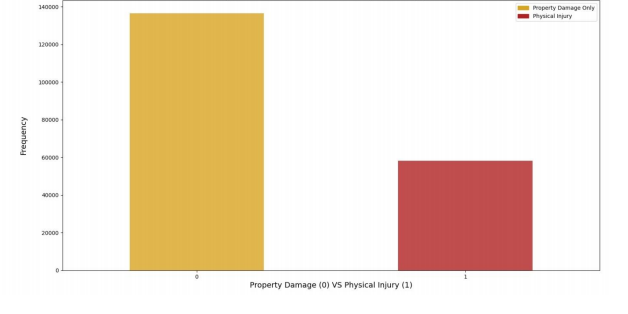
|  |  |
| --- | --- |
| Feature Variable | Description |
| INATTENTIONIND | Whether or not the driver was inattentive (Y / N) |
| UNDERINFL | Whether or not the driver was under the influence of drinking (Y / N) |
| WEATHER | Weather condition during impact collision (Overcast / Rain / Clear) |
| ROADCOND | Road condition during impact collision (Wet / Dry / Unknown) |
| LIGHTCOND | Light condition during impact collision (Light On / Dark with light on) |
| SPEEDING | Whether the car was above the speed limit at the time of collision (Y / N) |

1. **Methodology**

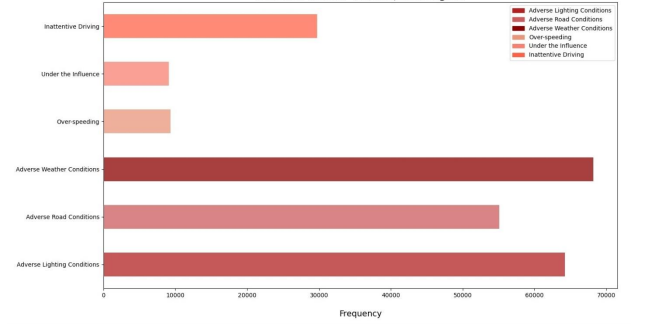
The dataset used for this project is based on car accidents which have taken place within the city of Seattle, Washington from the year 2004 to 2020. This data is regarding car accidents the severity of each car accidents along with the time and conditions under which each accident occurred. The data set used for this project can be found here.

**3.2 Exploratory Data Analysis**

Considering that the feature set and the target variable are categorical variables with the likes of weather, road condition and light condition being an above level 2 categorical variables whose values are limited and usually based on a particular finite group whose correlation might depict a different image then what it actually is. Generally, considering the effect of these variables in car accidents are important hence these variables were selected. A few pictorial depictions of the dataset were made in order to better understand the data.

**Accident Classification based on types**

The above figure illustrates, after data cleaning has taken place, the distribution of the target variables between Physical Injury and Property Damage Only. As it can be seen that the dataset is supervised but an unbalanced dataset where the distribution of the target variable is in almost 1:2 ratio in favour of property damage. It is very important to have a balanced dataset when using machine learning algorithms.

Classification of Accdient Causes

As mentioned earlier, a number ‘0’ as an element of an independent variable is supposed to depict the least probable cause of a severe accident. The graph above is supposed to depict all the nonzero values within each independent variable of the model and can be seen as the frequency of adverse conditions under which accidents took place. The factor which had most number of accidents under adverse conditions was adverse weather conditions while adverse lighting condition had the second most number of accidents caused by it. The factors which contributed the least to an instance of an accident are over-speeding and the driver being under the influence.

**3.3 Machine Learning Model Building**

The machine learning models used are Logistic Regression and Decision Tree Analysis. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. The Decision Tree Analysis breaks down a data set into smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. The reason why Decision Tree Analysis and Logistic Regression classification methods were chosen is because the Support Vector Machine (SVM) model is inaccurate for large data sets, while this data set has more than 180,000 rows filled with data. Furthermore, SVM works best with dataset filled with text and images.

1. **Results**

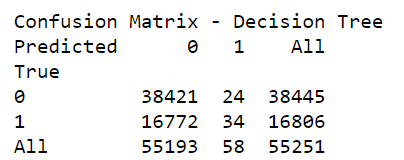
**4.1 Decision Tree Analysis**

Decision Tree Classifier from the scikit-learn library was used to run the Decision Tree Classification model on the Car Accident Severity data. The criterion chosen for the classifier was ‘entropy’ and the max depth was ‘6’.

**4.1.1 Classification Report**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1-Score |
| 0 | 1.00 | 0.70 | 0.82 |
| 1 | 0.00 | 0.59 | 0.00 |
| Accuracy |  |  | 0.70 |
| Macro Avg | 0.50 | 0.64 | 0.41 |
| Weighted Avg | 1.00 | 0.70 | 0.82 |

**4.1.2 Confusion Matrix**

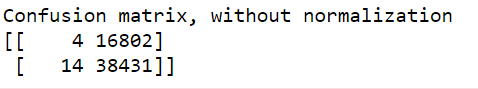
****

**4.2 Logistic Regression**

Logistic Regression from the scikit-learn library was used to run the Logistic Regression Classification model on the Car Accident Severity data. The C used for regularization strength was ‘0.01’ whereas the solver used was ‘bilinear’.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1-Score |
| 0 | 0.70 | 1.00 | 0.82 |
| 1 | 0.22 | 0.00 | 0.00 |
| Accuracy |  |  | 0.70 |
| Macro Avg | 0.46 | 0.50 | 0.41 |
| Weighted Avg | 0.55 | 0.70 | 0.57 |

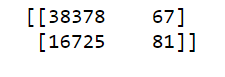
**4.2.1 Confusion Matrix**

****

**4.3 Advanced Gradient Boost**

The Advanced gradient boost from the gradient boost was used to run the advanced gradient boost model on the Car Accident Severity data. With random state of 20 and learning rate of about 0.1.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1-Score |
| 0 | 0.70 | 1.00 | 0.82 |
| 1 | 0.55 | 0.00 | 0.01 |
| Accuracy |  |  | 0.70 |
| Macro Avg | 0.62 | 0.50 | 0.42 |
| Weighted Avg | 0.65 | 0.70 | 0.57 |

**4.3.1 Confusion Matrix**

**5. 0 Discussion**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | **Average F1-Score** | **Property Damage (0) vs Injury (1)** | **Precision** | **Recall** |
| **Decision Tree** | 0.70 | 0 | 1.00 | 0.70 |
| 1 | 0.00 | 0.59 |
| **Logistic Regression** | 0.70 | 0 | 0.70 | 1.00 |
| 1 | 0.22 | 0.00 |
| **XGBoost** | 0.70 | 0 | 0.70 | 0.50 |
| 1 | 0.55 | 0.70 |

**5.1 Average F1-Score**

f1-score is a measure of accuracy of the model, which is the harmonic mean of the model’s precision and recall. Perfect precision and recall is shown by the f1-score as 1, which is the highest value for the f1-score, whereas the lowest possible value is 0 which means that either precision or recall is 0. The f1-score shown above is the average of the individual f1-scores of the two elements of the target variable i.e. Property Damage and Injury. When comparing the f1-scores of the three models, we can see that they all possess the same average F1-Score but a higher precision was obtained for decision tree and recall than the other two models. Whereas, the Decision Tree model’s f1-score has a score of 0.70. Lastly, the f1-score of the Logistic Regression is at 0.70 which can be considered as an above average score. However, the average f1-score doesn’t depict the true picture of the models accuracy because of the different precision and recall of the model for both the elements of the target variable. Hence, it is biased more towards the precision and recall of Property Damage due to its weightage in the model.

**5.2 Precision**

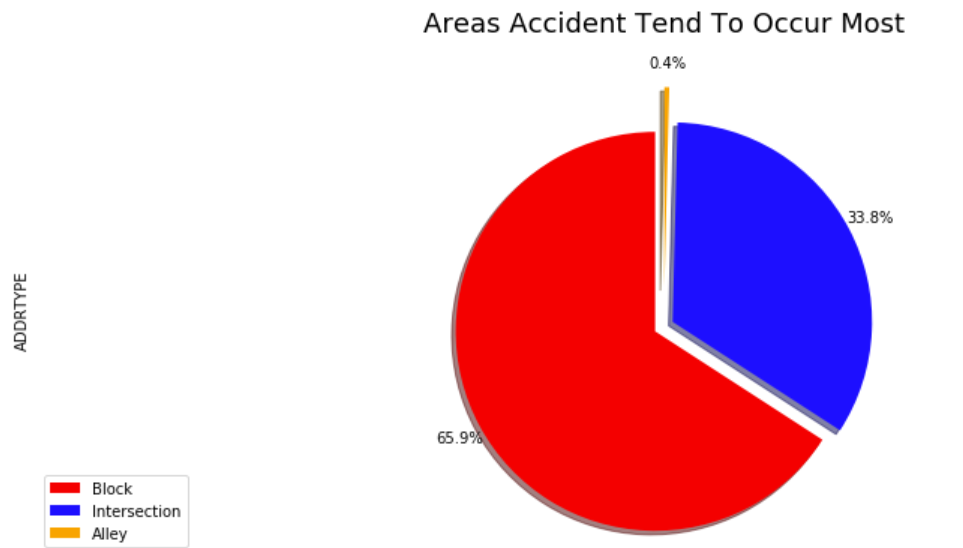
Precision refers to the percentage of results which are relevant, in simpler terms it can be seen as how many of the selected items from the model are relevant. Mathematically, it is calculated by dividing true positives by true positive and false positive. The precision for Property Damage is for Logistic Regression, whereas for Injury it is the Decision Tree. The Precision is calculated individually above in order to understand how accurate the model is at predicting Property Damage and Injury individually. For the Decision Tree the precision of 0 is 1.00 and for 1 it is 0.00 which is good. As for the Logistic Regression model, for 0 it is at 0.70 and for 1 it is 0.22. Lastly, for the Xgboost at 0 it is 0.70, which is highly accurate, however for 1 it is 0.55, extremely low. In terms of precision, the best performing model is the decision tree.

**5.3 Recall**

Recall refers to the percentage of total relevant results correctly classified by the algorithm. In simpler terms, it tells how many relevant items were selected. It is calculated by dividing true positives by true positive and false negative. The precision for 0 is 0.50when using the Xgboost model at 0.70 as for 1 it is the Logistic Regression model for 1.00 at 0 while 0 at 1.00. As for the Decision Tree, the recall for Property Damage is 0.70 and for Injury it is 0.59. The recall for Property Damage and Injury is the most balanced in terms of being good for both the outputs of the target variable.

1. **Conclusion**

When comparing all the models by their f1-scores, Precision and Recall, we can have a clearer picture in terms of the accuracy of the three models individually as a whole and how well they perform for each output of the target variable. When comparing these scores, we can see that the f1-score is highest for Xgboost at 0.70. However, later when we compare the precision and recall for each of the model, we can see that the Xgboost model performs poorly in the precision of 1 at 0.55. The variance is too high for the model to be selected as a viable option. When looking at the other two models, we can see that the Decision Tree has a more balanced precision for 0 and 1. Whereas, the Logistic Regression is more balanced when it comes to recall of 0 and 1. Furthermore, the average f1-score of the two models are very close but for the Logistic Regression it is higher by 0.70. It can be concluded that the both the models can be used side by side for the best performance. In retrospect, when comparing these scores to the benchmarks within the industry, it can be seen that they perform well but not as good as the benchmarks. These models could have performed better if a few more things were present and possible.

* A balanced dataset for the target variable.
* More instances recorded of all the accidents taken place
* Less missing values within the dataset for variables such as Speeding and Under the influence.
* More factors, such as precautionary measures taken when driving, etc. There should be a launch of developmental project in areas of severe accidents to minimize the effects of these factors. According to the diagram shown below

1. **References**

<https://www.asirt.org/safe-travel/road-safety-facts/>

<https://www.macrotrends.net/cities/23140/seattle/population#:~:text=The%20current%20metro%20area%20population,a%201.2%25%20increase%20from%202017.>

<https://www.seattletimes.com/seattle-news/data/housing-cars-or-housing-people-debate-rages-as-number-of-cars-in-seattle-hits-new-high/#:~:text=As%20of%202016%2C%20the%20total,are%20the%20number%20of%20cars.>

<https://wsdot.wa.gov/>

<https://www.nhtsa.gov/>