Final Report

Car Accident Severity Analysis

(Data Science Coursera Capstone Project)

Prepared By:

Rajkumar Bathani

1. Introduction

1.1 Background

National Highway Traffic Safety Administration of the USA suggests that the economical and societal harm from car accidents can cost up to *$871 billion* in a single year. Seattle, a city on Puget Sound in the Pacific Northwest, is surrounded by water, mountains and contain thousands of acres of parkland. Washington State's largest city, it's home to a large tech industry, with Microsoft and Amazon headquartered in its metropolitan area. There was a significant increase in the total number of vehicles in 2016 than in 2010. An increase in number of vehicles and car population rates can lead to higher numbers of accidents on the road.

1.2 Problem

The given dataset consists of 37 attributes that is associated with the car accident happened in Seattle, Washington, in which some or all can be used to train the Machine Learning model. Data that might contribute in predicting accident severity, Now, it would be great if you were warned of weather conditions and the road conditions, and about the possibility of you getting into a car accident and how severe it would be, So that you would drive more carefully, or even change your travel if you are able to.

2. Business Problem

Due to the increasing number of accidents and number of fatalities and damage, National Highway Safety Administration of the USA suggests that the economical and societal harm from car accidents can cost up to $871 billion in a single year. Apart from that huge number of fatalities has been recorded due to the accidents. The project aims to predict how severity of accidents can be reduced based on a few factors. The reduction in severity of accidents can be beneficial to the Public Development Authority of Seattle which works towards improving those road factors and the car drivers themselves who may take precaution to reduce the severity of accidents.

2.1 Data Science Methodology to solve a business problem

The CRISP-DM methodology is a process aimed at increasing the use of data mining over a wide variety of business applications and industries. The intent is to take case specific scenarios and general behaviors to make them domain neutral. CRISP-DM is comprised of six steps with an entity that must be implemented in order to have a reasonable chance of success. The six steps are shown in the Figure 1. This section will associate each stages of CRISP-DM methodology with the problem of predicting car accident severity.

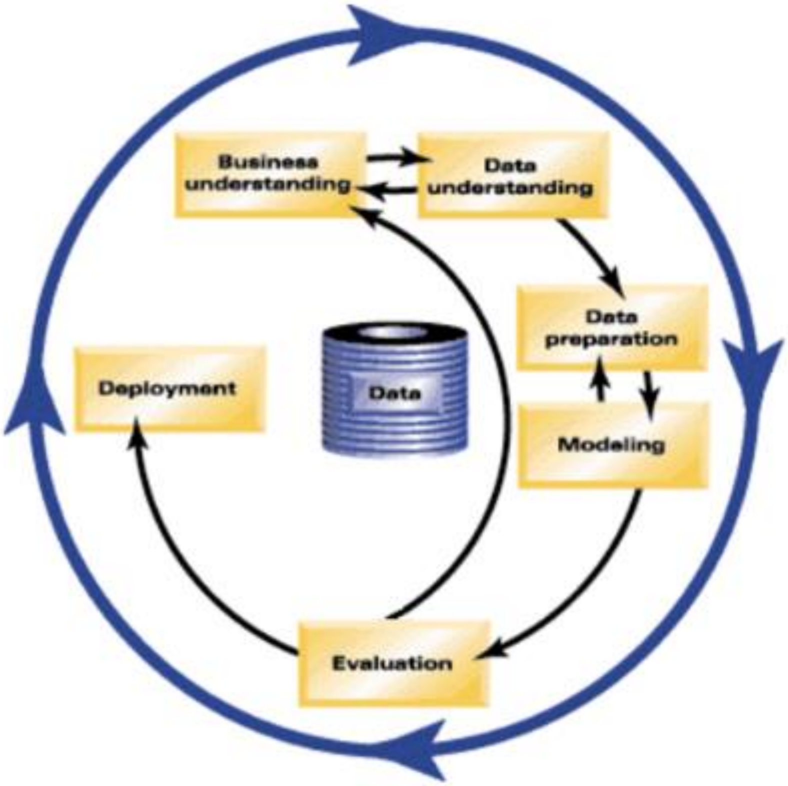


Figure 1: Data Science Methodology

The Business Understanding stage is the most important because this is where the intention of the project is outlined, the team of data scientists would communicate with the stakeholders and clarify the requirements. Stakeholders might have different objectives, biases, and modalities for relating information. Data Scientist will understand that how predicting the accident severity help the local city Police and the corresponding departments.

The Data Understanding relies on business understanding. Data is collected at this stage of the process. Once the goals are estimated after communication with stakeholders, the team of data scientists figure out the understanding of what the business wants and needs will determine what data is collected, from what sources, and by what methods. This combines the stages of Data Requirements, Data Collection, and Data Understanding from the Foundational Methodology outline.

Data Preparation once the data has been collected, it must be transformed into a useable subset unless it is determined that more data is needed. Once a dataset is chosen, it must then be checked for questionable, missing, or ambiguous cases. Many descriptive statistical tests are performed at this stage to obtain the statistical information. Also, exploratory data analysis is the integral part of data preparation process, where visualization plot and feature engineering might play a vital role in exploring the data and give business insights.

Modeling Once prepared for use, the data must be expressed through whatever appropriate models, give meaningful insights, and hopefully new knowledge. This is the purpose of data mining: to create knowledge information that has meaning and utility. The use of models reveals patterns and structures within the data that provide insight into the features of interest. Models are selected on a portion of the data and adjustments are made if necessary. Data Scientist will determine what approach to use for modelling (supervised or unsupervised based on the dataset).

Evaluation The selected model must be tested. This is usually done by having a pre-selected test, set to run the trained model on. This will allow you to see the effectiveness of the model on a set it sees as new. Results from this are used to determine efficacy of the model and foreshadows its role in the next and final stage.

Deployment In the deployment step, the model is used on new data outside of the scope of the dataset and by new stakeholders. The new interactions at this phase might reveal the new variables and needs for the dataset and model. These new challenges could initiate revision of either business needs and actions, or the model and data, or both.

This is how the CRISP-DM methodology can be applied on real life problems and each stages of this methodology helps in solving the problems step by step.

3. Understanding Data

3.1 Data Sources

The dataset used for this project is about car accidents which have happened within the city of Seattle, Washington from the year 2004 to 2020. This data is regarding the severity of car accidents, the severity of each car accidents along with the time and conditions under which each accident occurred.

3.2 Data Cleaning

The dataset has many null values for several attributes. The target variable column 'Severity Code' has values categorized into 1 and 2, that needs to be transformed into 0 and 1 where 0 indicates 'Property damage only' and 1 represents 'Injury Collision', in order to avoid ambiguities while model fitting and predictions. 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 represented by 0, and 1 represent Mushy and Wet was assigned to 2. If talking about Weather Condition, 0 represents Clear, 1 indicates Overcast, Windy is 2 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 an extreme level of accident severity. Whereas, there were unique data which is not preferred to remove them such as above-mentioned attributes with value like ‘Unknown’ and ‘Other’.

3.3 Feature Selection

The Severity Code is set as target variable and with respect to target variable other 6 features are selected such as; INATTENTIONIND, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, and SPEEDING. The below table describes how each feature contributes to the overall severity of accidents.

Table.1 Description of how each feature affects the severity of accident

Feature Variables	Description of Features
INATTENTIONIND	Whether or not the driver was inattentive (Yes/No)
UNDERINFL	Whether or not the driver was under the influence (Yes/No)
WEATHER	Weather condition during the time of accident (Rainy/Overcast/Snowy/Clear)
ROADCOND	Road condition during the time of accident (Wet/Dry)
LIGHTCOND	Light condition during accident (Lights On/Dark with light on)
SPEEDING	Was the car above the speed limit at the time of accident (Yes/No)

4. Exploratory Data Analysis

Exploratory Data Analysis is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task, as it helps in analyzing data sets to summarize their main characteristics, often the visual methods.

The below image shows that that target variable ‘SEVERITYCODE’ is not equally distributed, as it can be said that it is highly imbalanced distribution of attribute. The values such as ‘0’ represents ‘property damage’ and ‘1’ signifies ‘physical injury’. It is very important to have a balanced dataset, otherwise, the machine learning algorithm will produce biased results. Hence, SMOTE was used to balance the target variable in equally distributed sets.

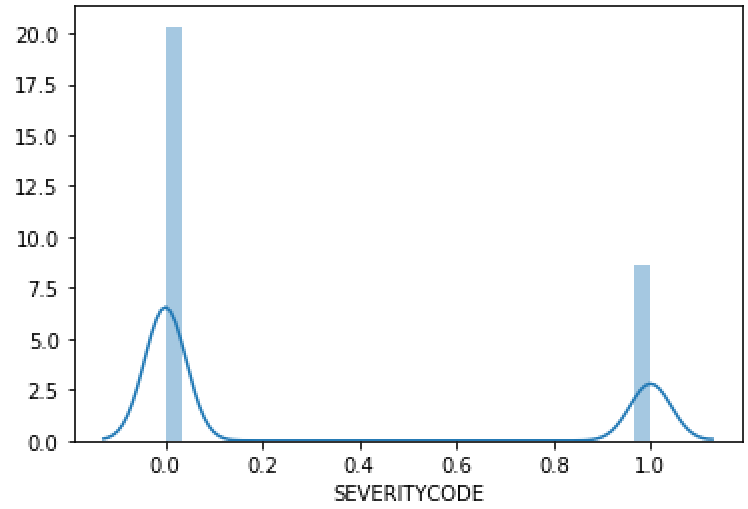


Figure 2 Distribution of accidents severity

The point where the accident happened also contribute to the severity of an accident. Accoding to Figure 3, there are three different values registered for the address type that shows the point where the accident happened such as; Block, Intersection and Alley. The counts for all the three different types of categories are as follows 126926, 65070, and 751 respectively.

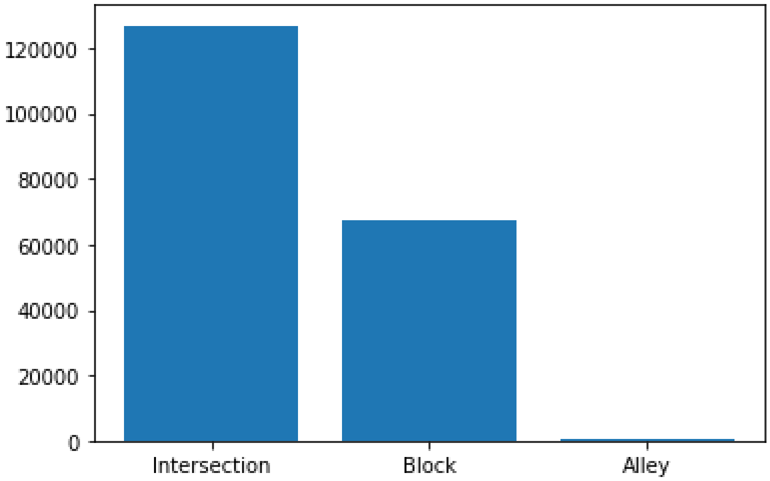


Figure 3 Area type of accidents

There is a need to encode all the selected features into a format which is interpretable by ML algorithm. For example, encoding weather conditions into numeric range of 0 to 3, where ‘0’ represents the best weather condition like ‘Clear’, and ‘1’ is for Overcast and Cloudy weather. ‘2’ is for Windy and ‘3’ is for Raining and Snowy conditions. All these encoding procedures should be applied to the selected features.

According to the below image, the independents variable which has contributed the most is adverse lighting condition and the variables such as under the influence and overspendings were the least contributors to the severity of an accident. This is how all features can be used to determine that which independent variable plays a vital road in causing the accidents.

5. Machine Learning Models

Different classification algorithms have been tuned and built for the prediction of the level of accident severity. These algorithms provided a supervised learning approach predicting with certain accuracy and computational time. These two properties have been compared in order to determine the best suited algorithm for this specific problem.

The used Machine Learning models are Logistic Regression and Decision Tree Classifier. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. The Decision Tree Classifier breaks down a data set into smaller subsets while at the same time an associated decision tree is incrementally developed.

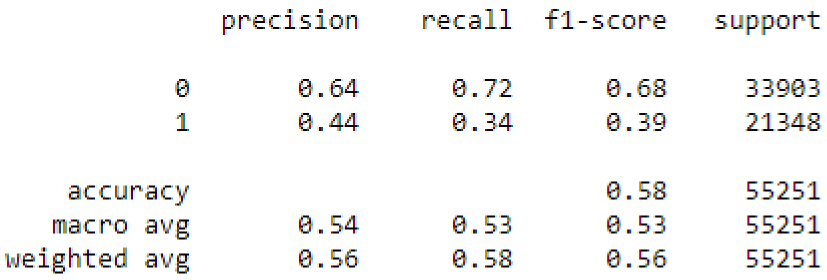
The Support Vector Machine (SVM) model is inaccurate for large data sets, while this data set has more than 180,000 rows filled with data.

6. Results

6.1 Decision Tree Classifier

Decision trees are built by splitting the training set into distinct nodes, where one node contains all or most of one category of the data. A decision tree can be constructed by considering the attributes one by one. First, choose an attribute from the dataset. Calculate the significance of the attribute in the splitting of the data. Next, split the data based on the value of the best attribute, then go to each branch and repeat it for the rest of the attributes. After building this tree, you can use it to predict the class of unknown cases.

What is important in making a decision tree, is to determine which attribute is the best or more predictive to split data based on the feature.



```
              precision    recall  f1-score   support

           0       0.64      0.72      0.68     33903
           1       0.44      0.34      0.39     21348

    accuracy                           0.58     55251
   macro avg       0.54      0.53      0.53     55251
weighted avg       0.56      0.58      0.56     55251
```

Figure 4: Classification Report of DTC

The above figure shows the classification report for the Decision Tree Classifier.

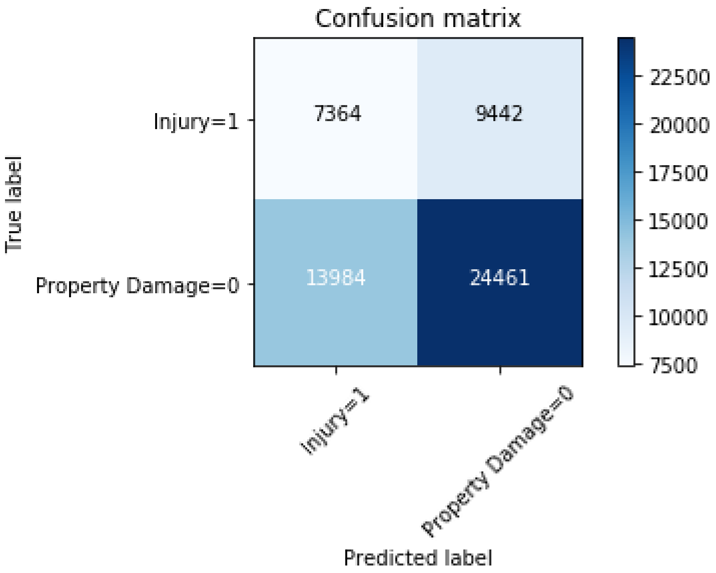
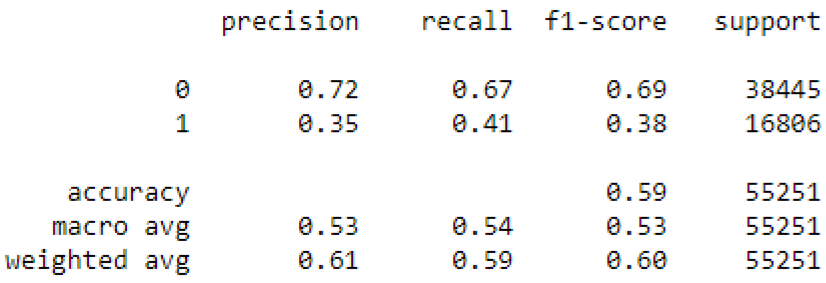


Figure 5 Confusion Matrix

6.2 Logistic Regression

Logistic regression is a statistical and machine learning technique for classifying records of a dataset based on the values of the input fields. Logistic regression is analogous to linear regression but tries to predict a categorical or discrete target field instead of a numeric one. In linear regression, we might try to predict a continuous value of variables such as the price of a house, blood pressure of a patient, or fuel consumption of a car. But in logistic regression, we predict a variable which is binary such as yes/no, true/false, successful or not successful, pregnant/not pregnant, and so on, all of which can be coded as zero or one.

In logistic regression independent variables should be continuous. If categorical, they should be dummy, or indicator coded. This means there is a need to transform them to some continuous value. Logistic regression can be used for both binary classification and multi-class classification.



```
              precision    recall  f1-score   support

           0       0.72      0.67      0.69     38445
           1       0.35      0.41      0.38     16806

    accuracy                           0.59     55251
   macro avg       0.53      0.54      0.53     55251
weighted avg       0.61      0.59      0.60     55251
```

Figure 6 Classification Report of Logistic Regression

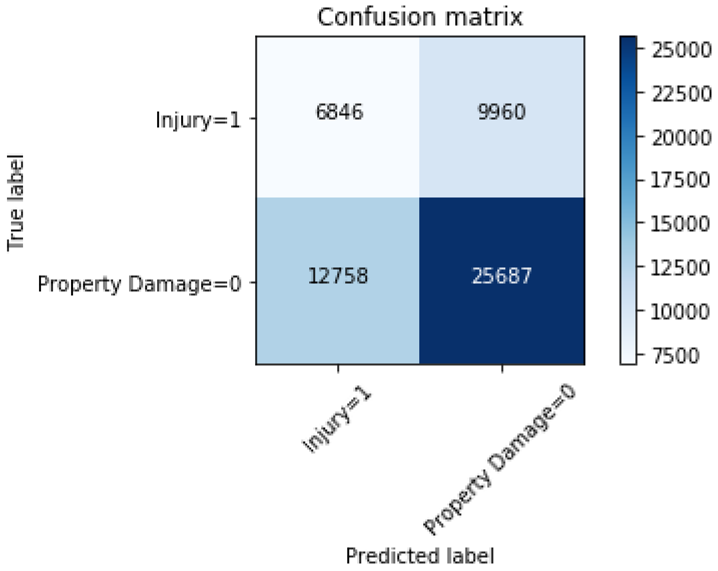


Figure 7 Confusion Metrix

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 are 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.

If making comparison among both models, the Decision Tree model's F1-score is the lowest of 0.56. Lastly, the f1-score of the Logistic Regression is at 0.60 which can be considered as an above average score. However, the average f1-score doesn't depict the true picture of the model's accuracy because 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.

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 highest precision for Property damage is for Logistic Regression, whereas for Injury it is the Decision Tree.

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. sAs for the Logistic Regression, the recall for Property Damage is 0.67 and for Injury it is 0.41. The recall for Property Damage and Injury is the most balanced in terms of being good for both the outputs of the target variable.

7. Conclusion

This case study shows that the relationship between severity of an accident and some characteristics which describe the situation that involved the accident such as Weather Conditions, Road Conditions, Light Condition, Influence and attentiveness.

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 k-Nearest Neighbor at 0.75. When looking at 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 0.04. It can be concluded that the both the models can be used side by side for the best performance.

8. Recommendations

After reviewing the obtained results from the Machine Learning models, it can be recommended that, the developmental body for the city of Seattle can assess how many of these accidents have taken place on a place where road or light conditions were not ideal for that specific area and could launch development projects for those areas where most severe accidents take place in order to minimize the effects of these two factors. Whereas, the car drivers could also us this data to assess when to take extra precautions on the road under the given circumstances of light condition, weather condition, road condition, in order to avoid a severe accident.