Accident Severity Prediction in Seattle, Washington

Using Decision Tree, Logistic Regression, Naïve Bayes & Linear Discriminant Analysis

IBM Data Science Professional Capstone
Coursera

September 23, 2020

By: Dolu Okeowo

Table of Contents

Executive Summary

The objective of this project is to predict the severity of an accident in Seattle, Washington using machine learning models. Features such as road, weather and light condition were used for this prediction.

Predicting a categorical target data requires using a supervised classification. We used Decision Tree, Logistic Regression, Linear Discriminant, Naïve Bayes model and evaluated the result of the models using precision, recall, accuracy and F1 score. The results obtained using the different models were not significantly different. The accuracy of all models was about the same i.e. 0.5. We suggest training the model with more features to obtain a better accuracy in the future. We could also implement ensemble learning to further improve the predictability of the model.

Introduction

Every day, millions of people commute using either their personal cars or public transportation for various reasons such as work, leisure etc. The urgency to get to our destination, lack of attention, driving under influence, road and weather condition, failure to obey traffic rules and regulations are one of the many reasons that causes accidents. This does not only lead to the loss of lives of the drivers and passengers involved, but this could also involve pedestrians and cyclist.

The benefits of being able to predict the severity of an accidents will not only forestall such mishaps but could help first responders send the right assistance to mitigate the damage and loss of lives of people. Emotional and socio-economic importance to the family, friends, and dependents of accident victims cannot be quantified.

In this project, we attempt to predict the severity of an accident using the machine learning algorithms such as Logistic regression, KNN, Linear Discriminant, and Naïve Bayes and Decision Tree. We used features such as road, light and weather condition. We evaluated the result of our prediction using F1 score, accuracy, precision and recall.

The prediction algorithm with not only help commuters determine whether it is safe to drive based on some certain conditions that are known in advance. Further, this model would also be beneficial to the first responders in Seattle, Washington in deploying the right personnel to accident scenes by knowing the severity of the accident ahead. Overall, this model could save lives and properties, and also reduce the cost of deploying by deploying the appropriate first responders.

Data

The accident data set of Seattle, Washington used in this project was provided by Coursera. The data consist of **37 attributes** both numerical and categorical, and **194,673 records**. The list below shows the 37 attributes provided in the data.

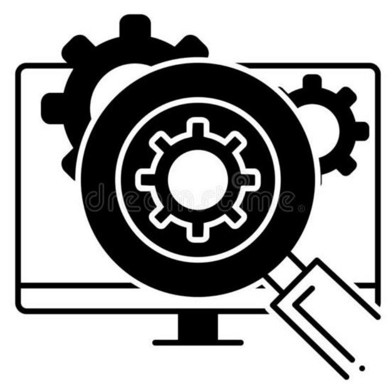
Data Columns: ['SEVERITYCODE', 'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO',
'STATUS', 'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE',
'EXCEPTRSNDESC', 'SEVERITYCODE.1', 'SEVERITYDESC', 'COLLISIONTYPE',
'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INCDATE',
'INCDTTM', 'JUNCTIONTYPE', 'SDOT_COLCODE', 'SDOT_COLDESC',
'INATTENTIONIND', 'UNDERINFL', '**WEATHER**', '**ROADCOND**', '**LIGHTCOND**',
'PEDROWNOTGRNT', 'SDOTCOLNUM', 'SPEEDING', 'ST_COLCODE', 'ST_COLDESC',
'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR']

For more details on the above attributes, see the metadata provided.

The **SEVERITYCODE** column is the target variable that contains the severity of the accident we are trying to predict. Based on this data, the severity can be categorized into bodily injury and property damage only collision.

To achieve the above goal, we select features that are known prior to the accident such as; **road, light and weather condition.**

Methodology



To predict the severity of accidents in Seattle, Washington, predicting a non-continuous variable, we implemented a supervised machine learning model. The target variable is the SEVERITY columns which consist of not sever and very severe.

Data Preprocessing & Cleaning

Data set usually do not come in the right format required before passing it into machine learning algorithms. We performed the following preprocessing and transformation on the dataset. Missing data, balancing, One-Hot encoding of categorical data.

Balance the data to prevent bias in our model

Missing Data

We dropped any row that contains at least one missing record. Prior to deletion of missing records, there was a total of 194,673 records, after deletion, there was a total of 189,337 records. This accounts of 6% of dataset deleted

Balancing imbalanced dataset

Figure 1 shows the imbalanced dataset of the target variables. Using this data without resampling the data will introduce bias into our model prediction.

In order to balance this data set, we used the Synthetic Minority Over-sampling **(**SMOTE) oversampling technique in the imblearn library. Figure 2 shows the result of the dataset after oversampling

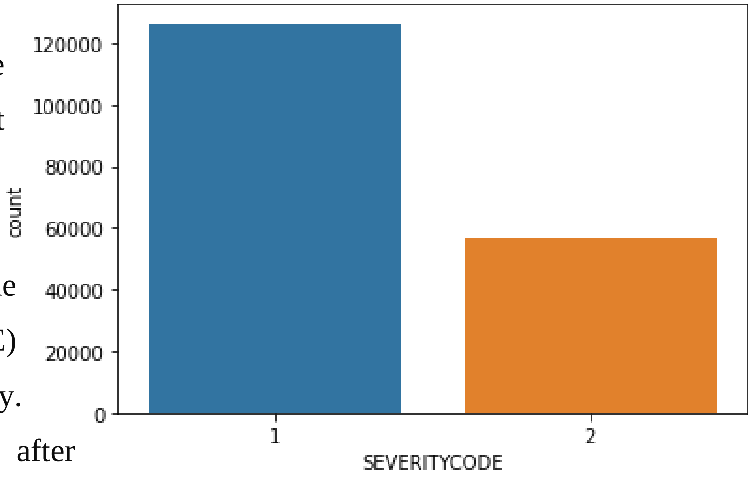


Figure 1: Imbalance Target Data

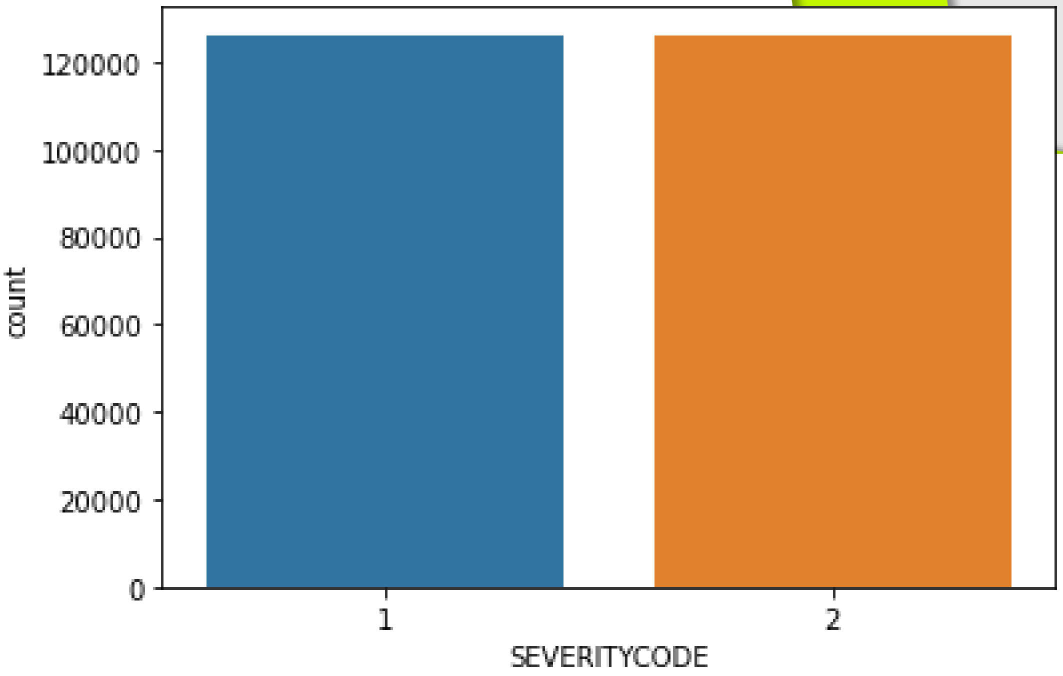


Figure 2: Distribution of target column after oversampling using the SMOTE technique

One-Hot Encoding:

The features available to model the accident severity are categorical data. Hence, it is important to convert the categorical data to binary data using One-Hot encoding. This method was chosen over the label encoder since our categorical data is not nominal. We used One-Hot encoding for the weather, light and road condition categorical features.

Exploratory data analysis

From figure 3, we observe that more accidents occurred despite the clear weather. Perhaps, this is due to the fact that the traffic is higher during clear weather thereby increasing the chances of accidents occuring.

We also observe that for all weather conditions, there was more property damage than injury.

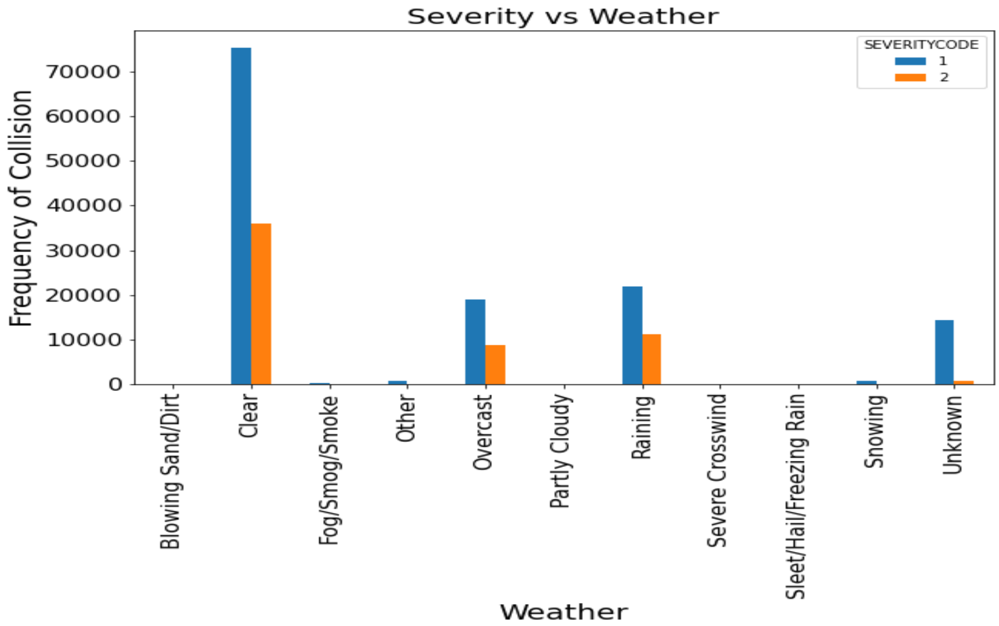


Figure 3: Histogram of Weather Count based on Severity of accident

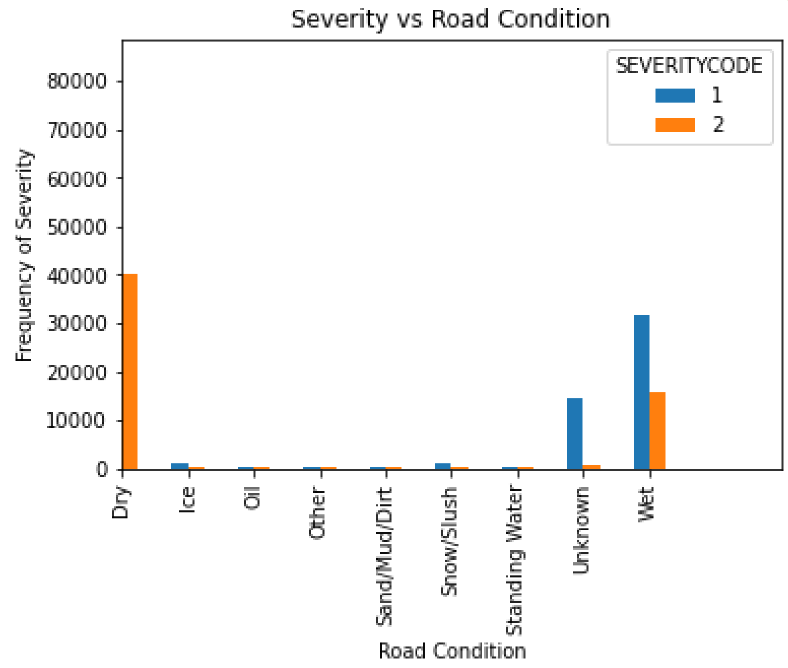


Figure 4: Histogram of Road Condition based on the Severity of accident

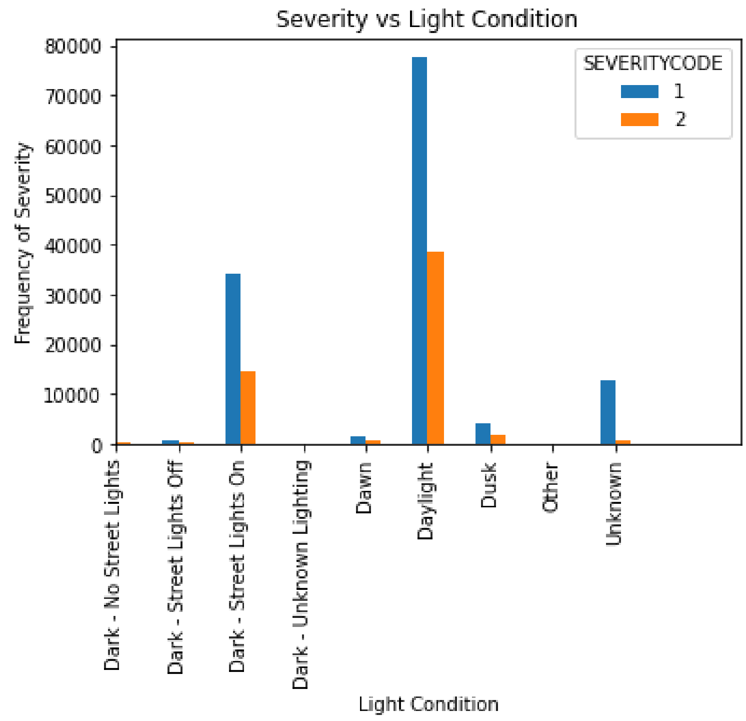


Figure 5: Histogram of Light-Condition based on the Severity of the accident

Train, Test Data

The data set was randomly split into *70% training and 30% test data*. The training data set was used to train the model while the test data set was used to evaluate the performance of the model.

Model selection

There are several supervised classification models that can be used to model the binary output of the severity of an accident. For the purpose of this project, we implemented: *Logistic Regression, Decision Tree, Linear Discriminant, and Naïve Bayes*.

Feature Selection

Selecting the right features to predict target variable is a very crucial step. One of the methods to select features is Sequential Backward Selection (SBC). In this process, features are sequential removed. However, this process could be computational intensive. Hence, for the simplicity of this project, we are selecting features that intuitively could affect the severity of an accident. The following features were selected to model the severity of accidents**: *WEATHER, ROADCOND, and LIGHTCOND***.

Cross Validation

One of the methods used in prevent overfitting or underfitting in machine learning is using a cross-validation method. In order to address this issue, we implemented cross validation using the K-fold of 5. Based on this method, a certain percentage of the data is held and used for cross validation of the model and the metric is evaluated. This allows each record to be used for both testing and training the data. At the end of this process, the average value of the metric is computed.

Result

To predict the severity of accident in Seattle, Washington, we used; decision tree, logistic regression, Naïve Bayes and linear discriminant analysis.

Table 1 shows the result of the metrics used to evaluate the model. To evaluate the model, we used the following metrics; F1 Score, precision, recall and accuracy. We observe from the table that the results obtained using the four different models was very similar. We obtained an average accuracy of 0.5 for all the models used.

Table 1: Evaluation of metrics for the algorithms

Models	F1 Score	Precision	Recall	Accuracy
Decision Tree	0.42	0.63	0.31	0.56
Logistic Regression	0.41	0.63	0.3	0.56
Naïve Bayes	0.66	0.5	0.98	0.5
Linear Discriminant Analysis	0.44	0.6	0.35	0.55

Discussion

The result obtained in this model was not as impressive as expected. To improve this model, we suggest adding more features to improve the accuracy of the model. Using 3 features; light, weather and road condition was not sufficient in training the model.

We are also optimistic that the model could also be improved through hyper-parameter tuning and evaluating our models for bias and overfitting. These recommendations should be implemented to make better predications for Seattle, Washington.

Furthermore, hot-spot analysis could also be done in the spatial domain and can be added to the model as an addition feature.

Conclusion

We can conclude from the results obtained from the metrics that predicting the severity of an accidents will require more features to make better prediction. Using different algorithms further confirms that the low accuracy obtained was not due to the choice of the algorithms. However, the model obtained could still be used to predict the severity of accidents.

In conclusion, motorist as well as first responders would be beneficiary to the accident model prediction tool when deployed.

Acknowledgement

Glory to God for the gift of life, from Whom all blessings flow. Special thanks to my friend Tolulope Omotoso who motivated me to complete this capstone.