Predicting the Severity of Car Accident

Abiola Adeyinka

September 30, 2020

1. INTRODUCTION / BUSINESS PROBLEM

Traffic accidents are undesirable and unforeseen events that have negative consequences like injury, damage and even death which are mainly caused by human, vehicle, road and/or environmental factors. According to statistics from the World Health Organization (WHO), road traffic accidents resulted to the death of about 1.35 million people and between 20 – 50 million people are left injured or disabled annually (World Health Organization Road traffic injuries, 2020). It was also reported that approximately 90% of road traffic accidents causes socioeconomic loss to individuals, families and the nation.

The problem of traffic accident in the world affects people, properties and the nation as a whole, the impact of which leads to death or serious injury, destruction of property and economic loss. A successful solution would be to predict the severity of accidents and also to determine the major factors that leads to traffic accidents in order to readdress those impacts and prevent traffic accidents. This will help to provide useful information to **people, emergency respondents and governments** to estimate the possibility of an accident, evaluate the severity of accidents and implement accident prevention procedures.

Therefore, this work focuses on conducting an accident severity prediction model by employing classification modeling techniques like Naïve Bayes, Decision Tree, Random Forest then the accuracies of the models will be compared and the best model will be selected for accident prevention. Also, exploring the major factors that impact accident severity.

2. DATA UNDERSTANDING

Data for this study was collected by the Seattle Police Department and recorded by Traffic Records Group which consist of collisions (traffic accident) data from 2004 to present. The dataset is the Seattle vehicle collisions data. You can find the Dataset by Clicking Here.

The dataset contains 194,673 rows and 37 columns (features). The predictor or target variable will be 'SEVERITYCODE' because it is used to measure the severity of an accident. The feature description of the dataset can be found by Clicking Here.

We will use data science technique on the data by incorporating the collision data per collision, the weather condition, road condition and lighting condition at the time of crash. These will allow us to determine the correlation between the weather condition and total number of crash as well as the lighting condition based on the weather condition. We then determine the impact of weather condition and road condition on the severity of accident.

Example of the Dataset.

```
data = pd.read_csv('Data-Collisions.csv')
data.head()
```

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight

5 rows × 38 columns

3. METHODOLOGY

i) Data Preprocessing

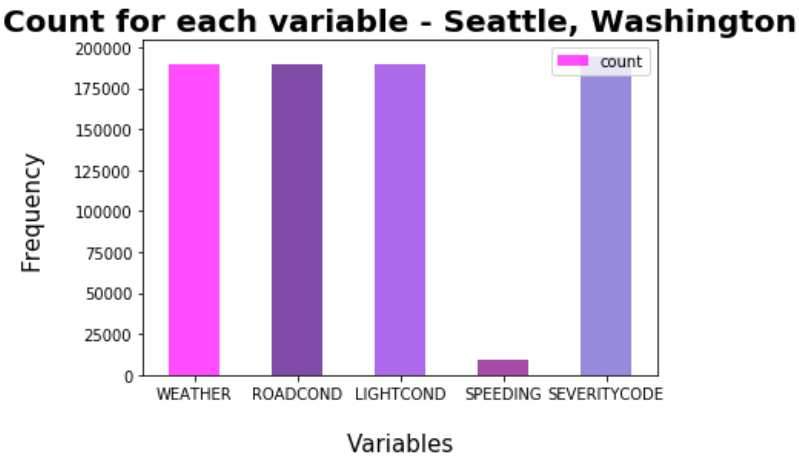
There are a lot of problem with the data set keeping in mind that this is a machine learning project which uses classification to predict a categorical variable. The dataset has total observations of 194673 with variation in number of observations for every feature. The total dataset was high variation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

The models aim was to predict the severity of an accident, considering that, the variable of severity code in the form of 1 and 2 which were label encoded to the form 0 and 1 respectively. Also, the variables for weather, road condition and lighting condition are converted from categorical features to numerical values.

ii) Feature Selection

A total of 3 features were selected to predict the target variable being severity code.

Feature	Description
ROADCOND	Road condition during collision (wet, dry)
LIGHTCOND	Light condition during collision (light on, dark …)
WEATHER	Weather condition during collision (clear, rain, overcast …)



iii) Exploratory Analysis

An exploratory analysis was performed in order to determine what type of methodology and machine learning will be most appropriate. From the exploratory analysis, it was seen that the dataset is supervised but imbalanced where the distribution of the target variable is in almost 1:2 ratio in favor of 1. It is very important to have a balanced dataset when using machine learning algorithms. Hence, an under sampling using the sklearn resample library was performed on the data to balance the target variable in equal proportions in order to have an unbiased classification model which is trained on equal instances of both the elements.

iv) Machine Learning Model Selection

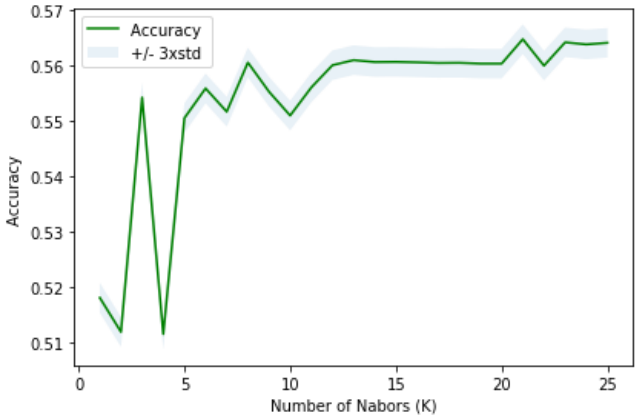
The classification models used are K-Nearest Neighbor, Decision Tree, Support vector machine and logistic regression. K-Nearest neighbor is a simple algorithm that stores all available cases and classifies new cases based on the similarity measure. Decision Tree analysis breaks down a data set into smaller subsets while at the same time an associated decision tree is incrementally developed. Logistic regression is a statistical model that in its basic form uses logistic function to model a binary dependent variable. Support vector machine was not used because the model is inaccurate for large data set and works best with data filled with text and images.

4. RESULTS

K-Nearest Neighbor

K-Nearest Neighbor classifier was used from the scikit-learn library was used to run the classification model on the data collision dataset. The best k, as shown below, used for the classifier was the highest elbow bend exists at 21. The balanced data was used to fit and predict the target variable.

Best KNN value



Decision Tree Analysis

Decision tree classifier from the scikit-learn library was used to run the classification model on the data collision dataset. The criterion used for the classifier was “entropy” and the max depth was “7”. The balanced data was used to fit and predict the target variable.

Logistic Regression

Logistic Regression classifier from the scikit-learn library was used to run the classification model on the data collision dataset. The C used for the regularization strength was “6”. The balanced data was used to fit and predict the target variable.

5. DISCUSSION

Algorithm	Jaccard	F1-score	Log Loss
K-Nearest Neighbor	0.564	0.540	NA
Decision Tree	0.566	0.545	NA
Logistic Regression	0.526	0.511	0.684

Recommendation

After assessing the data and output of the machine learning models, recommendations can be made to the stakeholders. The car drivers could use the 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 severe accident. Also, the government could use the data to launch development projects in areas were accidents have been recurring due to the road and lighting conditions by trying to minimize the effects.

6. CONCLUSION

When comparing the three models by their f1-sores, Jaccard score and log loss score, 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. Based on the above report, Decision tree is the best model to predict car accident severity because it has the highest Jaccard and f1-score of 0.566 and 0.545 respectively. Based on the dataset provided for this capstone from weather, road, and light conditions pointing to certain classes, we can conclude that particular conditions have a somewhat impact on whether or not travel could result in property damage (class 1) or injury (class 2).

Further Recommendations

The models can be performed better if a few things are present.

- One-hot encoding can be tried for some of the features

- More instances recorded of all the accidents taken place
- More factors, etc.

References:

1. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries
2. https://wsdot.wa.gov/