

Predicting the Severity of Collision

Doddi Durga Prasanna
September 2020

1. Introduction

The number of vehicular accidents occurring across the world are increasing day by day. Motor vehicle accidents was one of the leading cause of deaths in US in recent times. Careless driving habits augmented with drinking or under-age driving were found one of the leading factors for the vehicular accidents. Public monitoring using CCTV footage and application of complex computer vision techniques helped civil personnel in efficient monitoring and surveillance of US roads. But still the problem persists because of lack of resources with US civil bodies and careless attitudes of US citizens.

Problem

The number of vehicular accidents in US are increasing each day. It is becoming increasingly difficult for govt authorities to identify, track, and prevent the accidents at a granular level. Identifying the severity of the damage is a manual process, where a person/official analyses the accident zone physically present and classifies it for further down-stream processes.

Interest

Given the historical accident information available at a granular level (explained in next slide), we would like to use artificial intelligence to predict the severity of the damage occurred. This information will be utilized downstream for many processes such as –




Stakeholder	Use case	Benefit
 Insurance agencies	Accidental/damage information is quintessential for insurance agencies to accurately process claims in very short time	<ul style="list-style-type: none">• Faster claims resolution• Low loss/frauds in claims• Better customer Exp
 Govt Agencies (Police)	Govt agencies would need the data for social monitoring activities and targeted policing. Efficient utilization of Govt resources lead to less crime/accidents.	<ul style="list-style-type: none">• Efficient utilization of resources• Low crime rate• Increased social monitoring
 Mass Public Apps (GMaps)	Augmenting the zonal accidental data to apps like GMaps will help the mass public to drive safely and increased awareness towards safety	<ul style="list-style-type: none">• Reduced accidents• Increased public awareness• Increased customer exp

Figure 1: Possible use cases and benefits of solving this problem

2. Data acquisition and cleaning

Data Acquisition

We have collected a sample data set of collisions happening in US from the year of 2004 – 2012. The data set is publicly scrapped by IBM team and can be found [here](#). The data set was provided by SPD and recorded by traffic records. This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. The metadata information for the above data set can be found [here](#).

Data Set Basics	
Title	Collisions—All Years
Abstract	All collisions provided by SPD and recorded by Traffic Records.
Description	This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present.
Update Frequency	Weekly
Data Set Size	194674 records
Features/Attributes	37
Feature types	Multi-modal (int, Float, String, Object, Data time, etc.)

Figure 2: Data set high level details

Data Cleaning

The downloaded data set is of humongous volume with approximately 1.9 lack records and 37 different attributes. The labels of the attributes are provided in all capital format with truncated strings. For better understanding, we have renamed the labels of the attributes to human readable formats. One initial good observation found was the data set is consistent in time frame, i.e. weekly data is provided without any lapse.

There were several problems with the data set considered. Firstly, the data set has significant number of missing values for few of the features/attributes. The missing values were treated in two ways – 1. Removed the entire entry since we have a large amount of data set available and the removed entry will not make any significant difference, and 2. If the missing value is of type Int/float, based on logical inferences, we have replaced the missing values with mean/median as applicable.

Secondly, there are two date fields available in the data set which are default considered as Object type by Pandas. The type casting methods are used to convert the attributes to date-time format without any loss of data. Several other fields are in-correctly read by Pandas with wrong types. Used type casting to convert the variables to desirables types for further processing.

Feature Selection

Post cleaning the data, the resultant data frame was analyzed using the domain knowledge for feature selection. The data contained a lot of features which are identifiers (unique) which can result in no help for our classification exercise. Hence, these were removed. There are few features which are state assigned codes for the collision type dropped because the collision severity variable can be better explained using other variables and these became

redundant. Few features/attributes are textual based descriptions which are out of scope for my considerations since deep learning based text analysis and classification was not my current objective. Features involving geometry fields/unique geometric identifiers were removed since they are not useful for better explanation of the dependent variable.

Feature Kept	Features Removed	Reasoning
SHAPE , ADDRTYPE , SEVERITYCODE, COLLISIONTYPE , PERSONCOUNT , PEDCOUNT , PEDCYLCOUNT , VEHCOUNT , INJURIES , SERIOUSINJURIES , FATALITIES , INCDTTM , JUNCTIONTYPE , INATTENTIONIND , UNDERINFL , WEATHER , ROADCOND , LIGHTCOND	OBJECTID, INCKEY, COLDETKEY, INTKEY, , SEGLANEKEY , CROSSWALKKEY	Unique identifiers for the row; not very useful
	EXCEPTRSNCODE, ST_COLCODE , SDOT_COLCODE , EXCEPTRSNCODE	Special unique codes which are better explained by other variables
	EXCEPTRSNDESC , LOCATION , SEVERITYDESC, SDOT_COLDESC, ST_COLDESC	Textual variables (text analysis out of scope)
	HITPARKEDCAR , SPEEDING, PEDROWNOTGRNT, SDOTCOLNUM, INCDATE	Y/N type of variables with Incomplete/large number of missing values

Figure 3: Feature selection criteria

3. Exploratory Data Analysis

Target variable distribution is calculated to understand if the data set is too biased or imbalanced. The variable histogram is plotted in the below figure. We can conclude that both unique values of the variable (1 & 2) have significant number of samples to be considered the data set as balanced. Further no enhancements or techniques are applied in this direction.

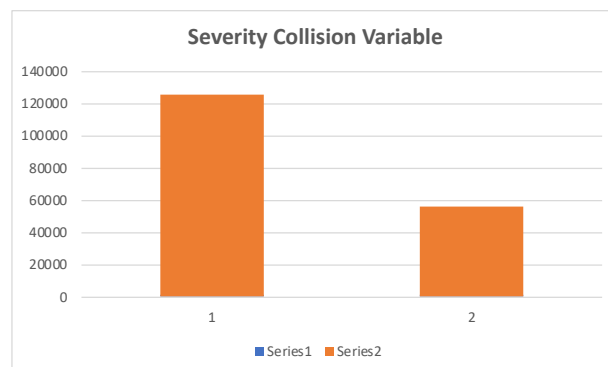


Figure 4: Dependent variable (Severity Collision) Distribution

Several categorical variables are plotted along the distribution of dependent variable to perform the exploratory data analysis. Below are the plots for reference.



Figure 5: Exploratory data analysis of categorical variables

Several observations and inferences were made looking at the above graphs. Address type variable is biased towards a single unique value, i.e. Alley type. Although there are no missing values at this stage these were considered in our analysis and are not omitted. Collision type variable also has the same problem with 3 unique values with low distribution. These are removed in the final consideration. Junction type and Weather variables have the problem of unknown values. After deliberation, we excluded these values since they might skew the final predictions.

4. Predictive Modelling

The dependent variable under consideration is a categorical variable with 2 unique values (1 & 2). Applying classification algorithms for prediction is the right course of modelling in this scenario. We have looked at the final pre-processed data set and converted all the categorical variables in the data sets to create dummies or duplicates. In this way, our final data set is homogeneous with Int/Float type and can be used for several Sci-Kit learn classification algorithms.

Test Train Split

Since we don't have any test data set readily available with the provided data set, we have to split our data set into 2. We have used Sci-Kit learn pre-processing library to divide the data set under consideration into Train set (80%) and test set (20%). The 80-20 rule is applied as a best practice and no further deliberation is done in this aspect.

Classification Modelling

We have considered using 4 classification models at hand to predict the collision severity. 1. KNN classifier, 2. Decision tree classifier, 3. SVM classifier, and 4. Logistic regression. Starting our modelling experimentation with KNN classifier, our objective was to find the best K with the highest accuracy. We have iterated the K values from 1-10, trained with the train data set and tested with the test set for accuracy calculation. We have found an inverted accuracy curve with the best accuracy observed at K=8.

Similarly, the decision tree classifier was trained using the Sci-Kit learn tree library. The criterion parameter was provided as 'entropy' which will maximize the information gain as we transcend the tree with max nodes as 4. We haven't considered experimenting with the

max nodes parameter since we observe as a best practice, 4 is ideal for decision tree modelling. The decision tree classifier is trained on the train data set and kept for evaluation.

SVM classifier was not optimized for the large data set under consideration. Even with heavy GPU (6GB VRAM) and RAM (16 GB), the performance of SVM classifier was sub-optimal. We had to abandon the training for the same reason. The logistic regression classifier was trained using Sci-Kit learn LR library on the train data set.

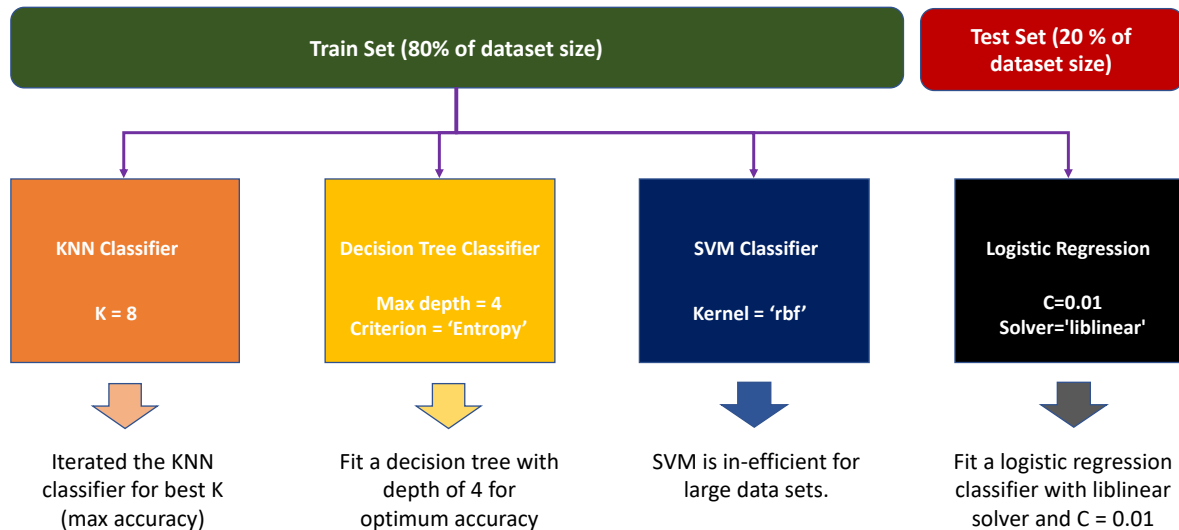


Figure 6: Data Modelling Overview

5. Model Evaluation and Conclusions

Model evaluation of all the 4 models was performed on the test data set which was kept isolated during the training/fit process. This will provide us the opportunity of identifying the best out-of-sample accuracy and removes the problem of over-fitting of the model. We have evaluated/calculated 3 key accuracy metrics for each of the classifier:

1. Jaccard similarity score
2. F1 score
3. Log loss

Below is the final evaluation score of the models -

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.718	0.689	NA
Decision Tree	0.721	0.652	NA
SVM *	NA	NA	NA
LogisticRegression	0.724	0.670	0.563

Figure 7: Accuracy metrics of 4 models

We have observed that the Jaccard scores of all the models under consideration are near same but Logistic regression gave a superior score. But when considering the F1-score, KNN classifier performed better. Our recommendation is to go ahead with the logistic regression since the data set volume is large and the dependent variable is binary which is the strong suite of logistic regression.

Conclusion

In this exercise, I have performed the prediction of collision severity of the accidents happening in US using diverse public data available. We have also seen that the trained model is performing with decent out of sample accuracy of 0.724 on the data set. There are a lot of avenues of improvement in the data gathering and modelling aspects. These systems will have direct downstream benefits for diverse set of stake holders such as insurance agencies, Govt agencies, or individual public with multi faced benefits.

Future Avenues

We have highlighted 4 avenues of improvements/future scope below:

- Plug in with other domain specific data sets for more contextualization and accuracy
 - Insurance claims, user traffic behavior, etc.
- Accuracy of the models can be improved with larger curated data sets
- Multi-modal data inclusion will help the prediction more accurate
 - CCTV footage (image/video analysis)
 - Demographic data of the persons involved (age, sex, race, ethnicity etc.)