Predicting the Severity of Car Accidents

John Lim

October 07, 2020

1. Introduction

1.1 Background

Traffic accident numbers have inevitably been on the rise, due to population growth and an increasing number of vehicles on the road. In turn, this translates to challenges faced by the trauma centres in hospitals and the traffic police departments. Due to manpower constraints and logistics concerns, the optimal deployment of resources in quick time is largely dependent on the severity of the accident. Hence, having an accurate prediction of accident severity in quick time would have huge benefits.

1.2 Problem

This project aims to identify a highly accurate classifier that predicts the severity of car accidents in Seattle, USA, using only input features that were easily identified without much investigation on the accident site.

1.3 Interest

This project might be of interest to trauma centres in hospitals and traffic police enforcement. It may also provide some insight to researchers and policy makers within government bodies such as the Healthcare Authority and Land Transport Authority.

2. Data Source

Data for all traffic accidents that happened between Jan 2004 and May 2020 were provided by the Seattle Traffic Police Division. This includes all types of collisions. Each collision record has been given a Severity Code label (1 - Property Damage Only Collision, 2 – Injury Collision), which is the target our model will try to predict. The data also includes many attributes that might be useful inputs for prediction (see Table 1).

Table 1: Candidate predictors for car accident severity

Attribute	Description
ADDRTYPE	Collision address type: • Alley • Block • Intersection
COLLISIONTYPE	Collision type
PERSONCOUNT	The total number of people involved in the collision
PEDCOUNT	The number of pedestrians involved in the collision
PEDCYLCOUNT	The number of bicycles involved in the collision
VEHCOUNT	The number of vehicles involved in the collision
INCDTTM	The date and time of the incident
JUNCTIONTYPE	Category of junction at which collision took place
SDOT_COLCODE	A code given to the collision by Seattle Department of Transportation
INATTENTIONIND	Whether or not collision was due to inattention
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol
WEATHER	A description of the weather conditions during the time of the collision
ROADCOND	The condition of the road during the collision

3. Methodology

3.1 Preliminary feature engineering and statistical inference

The time of accident was extracted from the INCDTTM variable and categorized into four buckets ('midnight to 6am', '6am to 12noon', '12noon to 6pm', '6pm to midnight').

Univariate statistical testing was conducted, where each candidate predictor variable was compared between both groups of the target variable (SEVERITYCODE = 1 and SEVERITYCODE = 2).

For the numerical variables (PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT), the Wilcoxon rank sum test was used. For all other variables which are categorical, the Chi-squared test of independence was used. These preliminary statistical tests might provide clues as to which variables significantly affect car accident severity. For instance, if the resultant p-value from such a test is greater than, say 0.20, then it might be justifiable to not include it as a feature in the prediction model.

3.2 Missing data

All statistically significant variables determined from the preliminary testing were listed down for inclusion in the prediction model. Records having missing information for any of these variables were dropped.

3.3 Machine learning models

For the classification problem at hand, three models were compared, namely the support vector machine, logistic regression and random forest classifiers. These were deemed more suitable than kNN (due to presence of many categorical features) or Naïve Bayes (due to presence of numerical features).

The data was randomly split to form a training set (70% of samples), a validation set (15%) and a test set (15%). Hyperparameter tuning was done by estimating model parameters (given the hyperparameter value) on the training set, and then observing an accuracy measure on the validation set. The hyperparameter would be chosen such that it maximized the accuracy

measure. Due to the unbalanced levels of the target variable, we used F1 score as the accuracy measure. After optimal hyperparameters were chosen, we combined the training and validation sets to re-train the models, and then evaluated them on the test set.

4. Results

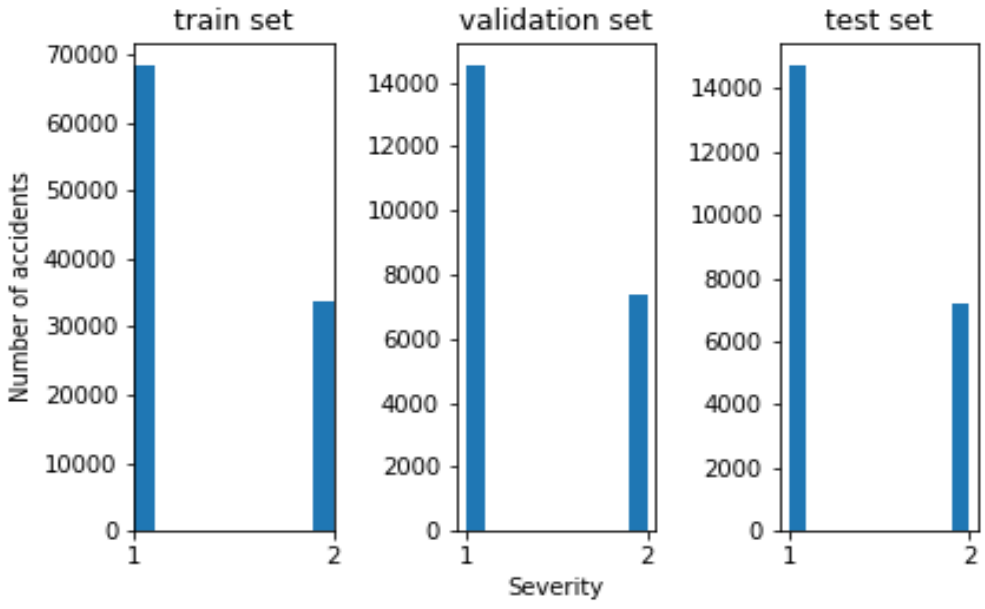
Univariate statistical test results are shown in Table 2 below. At this point, none should be excluded from the predictive model.

Table 2: Univariate tests for association with car accident severity

Variable	DataType	StatisticalTest	Pvalue
ADDRTYPE	object	Chi-squared test of independence	<0.001
COLLISIONTYPE	object	Chi-squared test of independence	<0.001
PERSONCOUNT	int64	Wilcoxon rank sum test	<0.001
PEDCOUNT	int64	Wilcoxon rank sum test	<0.001
PEDCYLCOUNT	int64	Wilcoxon rank sum test	<0.001
VEHCOUNT	int64	Wilcoxon rank sum test	<0.001
PERIOD	object	Chi-squared test of independence	<0.001
JUNCTIONTYPE	object	Chi-squared test of independence	<0.001
STRIKER	object	Chi-squared test of independence	<0.001
UNDERINFL	object	Chi-squared test of independence	<0.001
WEATHER	object	Chi-squared test of independence	<0.001
ROADCOND	object	Chi-squared test of independence	<0.001

Car accident severity distributions were similar amongst the training, validation and test sets, as shown in Figure 1 below.

Figure 1: Car accident severity distributions



Evaluation results on the test set are shown in Table 3 below, where all three models seem to have very similar performance based on F1 score, though the random forest classifier ranks top.

Table 3: Test set evaluation results

Classifier	F1 score
Random Forest	0.696
Logistic Regression	0.694
Support Vector Machine	0.687

5. Discussion and Conclusion

Simply predicting SEVERITYCODE = 1 for all accidents in the test set would have yielded an F1 score of 0.541. This shows that the classifiers all show significantly better capability in discriminating accident severity. In addition, the model still lacks certain information which would very likely be useful predictive features, such as:

- Speed of vehicles
- Whether any parts were detached / flew off from the vehicles
- Distance of detached parts from the accident site

Given the importance and motivation of accurately predicting the severity of accidents, it might be of interest to the authorities to explore ways to capture such features that might further increase the discriminative power of machine learning classifiers.