Capstone Project - Prediction of Car Accident Severity

By Yongyou Hu

1. Introduction: Business Problem

In the modern world, car accidents happen everyday and everywhere. Some accidents involve only minor or major property damages; however, others involve severe injuries or even fatalities. It is very important for all drivers to avoid accidents with high severity. This report aims to extract insights of accident data and predict accident severities based on several factors such as weather, light condition, location and so on.

The conclusion of insights will help drivers in the area make a correct judgement on the possibility of a severe accident involving injuries. It will also help drivers make a right decision if they should drive or not or if they should be more careful under certain conditions while driving. The targeted audience will be car drivers and traffic control department which routinely sends out traffic warnings and reminders.

2. Data Import and Cleaning

The dataset used by the report is from Seattle accident data downloaded from the course website (a file named "Data-Collisions.csv"). It has 194673 samples with 37 attributes (excluding the duplicate 'SEVERITYCODE'). The prediction will be 'SEVERITYCODE'.

2.1 Feature Selection

On examining the meaning of each feature among the 36 features, 7 of them were selected as features for modeling purpose. These features are 'VEHCOUNT', 'SPEEDING', 'ADDRTYPE', 'WEATHER', 'COLLISIONTYPE', 'PERSONCOUNT', 'INCDTTM'. However, 'INCDTTM' was translated into 'WEEKEND' and 'TIMEOFDAY'. So there are 8 features in total for the model. For modeling purpose, the following features were added with dummy ones.

"TIMEOFDAY": NIGHT, MORNING, AFTERNOON and EVENING

"ADDRTYPE": Alley, Block and Intersection

"COLLISIONTYPE": Angles, Cycles, Left Turn, Right Turn, Head on, Parked Car, Pedestrian, Rear Ended, Sidewipe and other

So there are 22 independent variables (features) in total for the model. Some features in the dataset were dropped because they are highly correlated with selected ones. For example, 'ROADCOND' is highly correlated to 'WEATHER' (Pearson coefficient > 0.9).

2.2 Data Cleaning

The dataset has many missing values which are not acceptable for modeling. For simplicity, those data samples with missing values were deleted. After cleaning, the total data samples were reduced from 194673 to 187755.

3. Methodology

In this project, my efforts were directly on predicting the car accident severities in the Seattle area based on several features. In the first step required data was collected and cleaned. Eight features (based on the meaning of each feature and correlations) were selected.

In the second step, relationship between major features and accident severity will be explored. Then various classification algorithms for machine learning will be investigated and compared for their performances.

In the final step the results will be discussed.

3.1 Exploratory Data Analysis

a. Relationship between TIMEOFDAY and SEVERITY

The accident date column, INCDTTM was separated into two features: TIMEOFDAY and WEEKEND to explore further detailed relationship between date and accident severity.

TIMEOFDAY	SEVERITYCODE	
AFTERNOON	1	45450
	2	21621
EVENING	1	26881
	2	11576
MORNING	1	28797
	2	12751
NIGHT	1	29705
	2	10974

Name: SEVERITYCODE, dtype: int64

TIMEOFDAY	SEVERITYCODE	
AFTERNOON	1	0.677640
	2	0.322360
EVENING	1	0.698988
	2	0.301012
MORNING	1	0.693102
	2	0.306898
NIGHT	1	0.730229
	2	0.269771

Name: SEVERITYCODE, dtype: float64

As shown above, AFTERNOON has the most accidents and highest ratio (32.2%) of Severity 2 accidents. This is likely due to highest traffic in the afternoon.

b. Relationship between WEATHER and SEVERITY

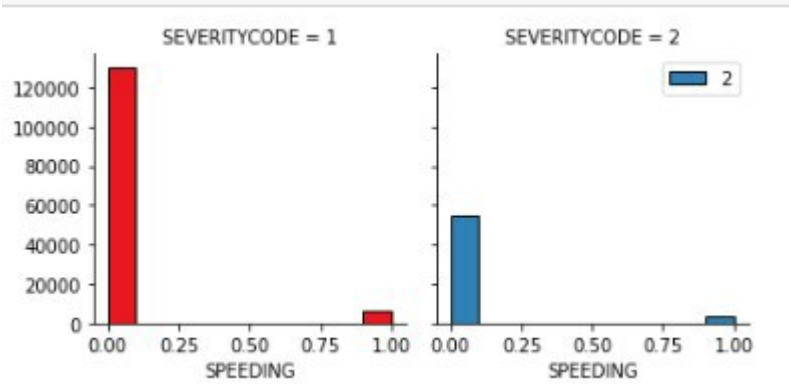
It is well known that bad weather or road conditions will cause more accidents with higher severity. However, the data below shows no big difference. The weather conditions don't affect accident severities a lot even though 'Raining' has a slightly higher number of Severity 2 accidents (33.7% vs. 32.2% in 'Clear'). It is possible that drivers were already aware of the bad weather and drove more safely. Also the road condition is highly correlated with weather condition.

WEATHER	SEVERITYCODE	
Blowing Sand/Dirt	1	0.732143
	2	0.267857
Clear	1	0.677509
	2	0.322491
Fog/Smog/Smoke	1	0.671353
	2	0.328647
Other	1	0.860577
	2	0.139423
Overcast	1	0.684456
	2	0.315544
Partly Cloudy	2	0.600000
	1	0.400000
Raining	1	0.662815
	2	0.337185
Severe Crosswind	1	0.720000
	2	0.280000
Sleet/Hail/Freezing Rain	1	0.752212
	2	0.247788
Snowing	1	0.811466
	2	0.188534
Unknown	1	0.945928
	2	0.054072

Name: SEVERITYCODE, dtype: float64

c. Relationship between SPEEDING and SEVERITY

Typically speeding will lead to more accidents with higher severity. It is confirmed by the data below.



SPEEDING	SEVERITYCODE	
0	1	0.705099
	2	0.294901
1	1	0.621665
	2	0.378335

Name: SEVERITYCODE, dtype: float64

As shown above, most accidents happened without speeding but speeding will more likely cause a Severity 2 accident (38% vs. 29%).

d. Relationship between ADDRTYPE and SEVERITY

The locations of accidents will affect the accident severity too. As shown in the data below, intersection is the dangerous place where major accidents with injuries can happen. Typically people will likely intend to be speeding and have red light violations at intersection.

ADDRTYPE	SEVERITYCODE	
Alley	1	0.890812
	2	0.109188
Block	1	0.762885
	2	0.237115
Intersection	1	0.572476
	2	0.427524

Name: SEVERITYCODE, dtype: float64

e. Relationship between COLLISIONTYPE and SEVERITY

collision type will definitely affect the accident severity. As shown in the data below, any collision involving pedestrians and cyclists will be 88-90% Severity 2 accident, which makes sense. Head On and Rear Ended are also among the collision types with highest Severity 2 accidents followed by Left Turn and Angles.

COLLISIONTYPE	SEVERITYCODE	
Angles	1	0.606863
	2	0.393137
Cycles	2	0.876041
	1	0.123959
Head On	1	0.568171
	2	0.431829
Left Turn	1	0.604784
	2	0.395216
Other	1	0.741318
	2	0.258682
Parked Car	1	0.943334
	2	0.056666
Pedestrian	2	0.898241
	1	0.101759
Rear Ended	1	0.569102
	2	0.430898
Right Turn	1	0.793737
	2	0.206263
Sideswipe	1	0.865276

3.2 Classification Models

There are two categories (Severitycodes 1 and 2) for the accident severity (labeled output of the model). So this is a classification problem. In this project, Logistic Regression, K-Nearest Neighbor and Support Vector Machines algorithms will be selected for machine learning.

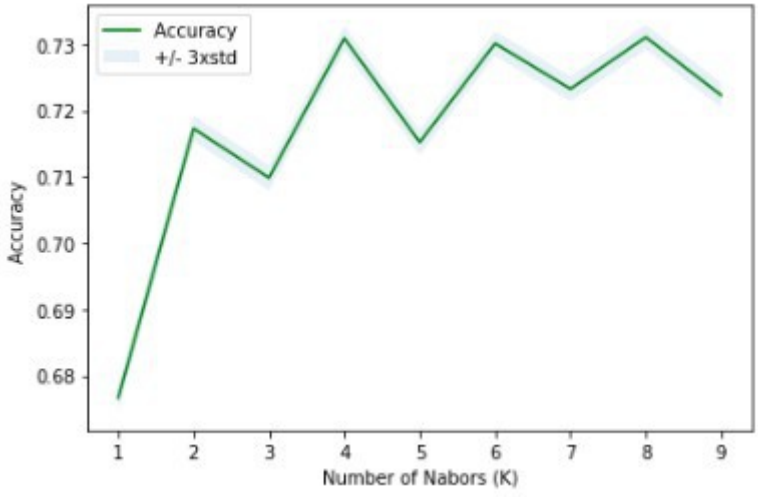
The dataset was split for 70%/30% for training and testing purpose for all models. The features were normalized and the labeled output was converted to 0/1 binary.

a. Logistic Regression

The optimal regularization parameter, C was found to be 0.1 based on a few trials. Then C = 0.1 was used for the final model. Solving this model was quite fast.

b. K-Nearest Neighbor (KNN)

The optimal neighbors, K was found to be 8 based on the curve below (refered to the highest accuracy). Then K = 8 was used for the final model. Solving this model was really slow.



c. Support Vector Machines (SVM)

The optimal kernel was found to be radial basis function (rbf) based on results of all

four kernels (linear, ploynomial, rbf and sigmoid). Then rbf was used for the final model. Solving this model was very slow too.

3.3 Model Performance

The table below shows performance comparisons among three models

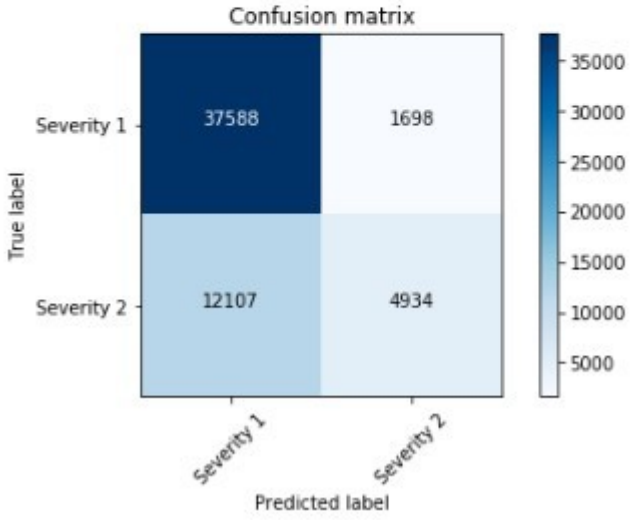
Algorithm Name	Jaccard	F1-Score	Log Loss
KNN	0.730	0.705	-
SVM	0.748	0.687	-
Logistic Regression	0.755	0.715	0.485

Per jaccard index and f1-score for model accuracy, the logistic regression algorithm has the best overall accuracy (a jaccard index of 0.755 and a f1-score of 0.715). More importantly, training KNN and SVM models are much slower than the logistic regression model. So the logistic regression model is used for this project.

It is important to check the precision and recall of Severity 2 accidents. Let us take a look at the confusion matrix of the logistic regression model.

	precision	recall	f1-score	support
0	0.76	0.96	0.84	39286
1	0.74	0.29	0.42	17041
accuracy			0.75	56327
macro avg	0.75	0.62	0.63	56327
weighted avg	0.75	0.75	0.72	56327

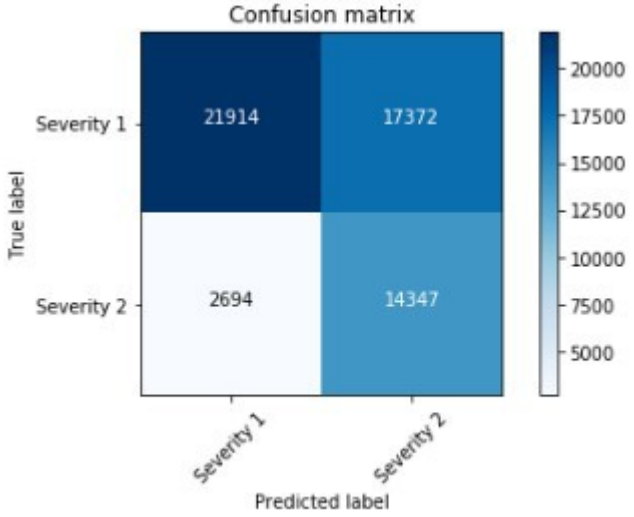
```
Confusion matrix, without normalization
[[37588  1698]
 [12107  4934]]
```



As shown above by the confusion matrix, the recall (0.29) and f1-score (0.42) for Severity 2 accident are too low, which is not helpful for avoiding injuries. One of the reasons could be unbalanced labels in the training sets (91547 Severity 1 vs. 39881 Severity 2). This can be improved by balancing the labels. A new training set with randomly selected 39881 samples of Severity 1 accident and 39881 Severity 2 was used to train the same model. The updated confustion matrix is as follows:

	precision	recall	f1-score	support
0	0.89	0.56	0.69	39286
1	0.45	0.84	0.59	17041
accuracy			0.64	56327
macro avg	0.67	0.70	0.64	56327
weighted avg	0.76	0.64	0.66	56327

```
Confusion matrix, without normalization
[[21914 17372]
 [ 2694 14347]]
```



Apparently, the recall and f1-score of Severity 2 accident have been greatly improved from 0.29 to 0.84 and 0.42 to 0.59, respectively. So balanced labels worked very well to improve the model.

4. Results and Discussion

In this study, the relationship between car accident severities and a couple of factors such as weather conditions, time of day and locations were analyzed. Classification models (Logistic Regression, KNN and SVM) were built to predict the car accident severities. Among three models, the logistic regression model is the fastest and most accurate one. This model can be very useful in helping car drivers and traffic control department in a number of ways. For example, it could help drivers decide if they should drive or not today based on weather conditions, number of passengers and time of day or when is the better time to drive.

The results show that weather and road conditions barely affected the labeled outputs because drivers were already aware of situations and drove more safely. But it doesn't mean people should ignore the bad weather and road conditions. It is well known that bad weather and road conditions can lead to more accidents if people ignore them. The traffic control department should still send out warnings and reminders when bad weather and road conditions exist. Most accidents including Severity 1 and Severity 2 happened in the afternoon when the traffic is usually heavy. Accidents due to speeding and involving pedestrians and cyclists are more likely Severity 2. So drivers should avoid speeding and drive more carefully when encountering pedestrians and cyclists. More attentions should be paid to intersections where more Severity 2 accidents happen.

It is important to use balanced labels to train the model, otherwise, the recall of Severitiy 2 accidents will be too low and won't help drivers avoid injuries.

5. Conclusion

Purpose of this project was to predict car accident severities based on driving conditions including weather, road, time of day and so on. It is concluded that drivers should avoid speeding and driving in afternoon if possible to mitigate the risk of injuries. When driving, they should be more careful at intersections and when pedestrians and cyclists are present.