The Prediction of car accident's severity

Introduction

Background

The US average number of car accidents is 6 millions every year. These car accidents resulted in more than 30000 fatalities and 3 million injuries annually. 60% of the those injuries resulted in permanent disabilities among the drivers. Although the number of fatalities per the total US population declined over the last two decades, the numbers started to increase in 2015 and continued to move up in 2016. According to business insider, there are seven common causes of car accidents in the US. Distracted driving due to paying attention to road, smartphones or applying makeup as well as driving under influence are the leading causes of car accidents. Breaking the speed limit is the second cause of car accidents. Although, all of the previous causes are driver related, there are other driver's unrelated causes such as bad weather, low visibility (night time), or road condition. To help the drivers to avoid being involved in a car accident, it will be important to predict how severe the accident would be if the driver got involved into an accident based on these factors causing accidents, so that the driver would drive more carefully or even change his travel if he is able to.

Problem

it is important to have data associate how severe the accident will be and some factors such as the weather condition, road status, and degree of visibility. With such data, the project aims to build a model that could predict the possibility of how sever the accident will be.

Interest

The car drivers will be interested in such model that predict the severity of the car accident based on the current circumstances of the driver and/or the trip , so that the driver would drive more carefully or even change his travel if he is able to.

Methodology

Data

The source of data:

The data is a shared collection of all car accidents occurred in Seattle city from 2004 to Present. This includes all types of collisions provided by SPD (Seattle Police Department) and recorded by Traffic Records.

Description of data:

The raw data is composed of 37 attributes that describe 194673 different accidents. By looking at these attributes, they can be categorized as shown in table 1.

Table 1: Attributes of the Traffic Records provided by SPD.

Category #	Category name	Attribute's name
1	Identification	OBJECTID, INCKEY, COLDETKEY,
2	Location	SHAPE, ADDRTYPE, INTKEY, LOCATION, CROSSWALKKEY
3	Time	INCDATE, INCDTTM,
4	Accident's description	SEVERITYCODE, SEVERITYDESC, COLLISIONTYPE, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, INJURIES, SERIOUSINJURIES, FATALITIES, JUNCTIONTYPE, SDOT_COLCODE, SDOT_COLDESC, SDOTCOLNUM, ST_COLCODE, ST_COLDESC, SEGLANEKEY, HITPARKEDCAR
5	driver's related causes of car accident,	INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SPEEDING
6	Environment's related causes of car accident.	WEATHER, ROADCOND, LIGHTCOND
7	Unknown	EXCEPTRSNCODE, EXCEPTRSNDESC

As shown in table 1, three of these attributes are just to give identification number for the accident. Four attributes give the location and the address type. Two features inform when the accident occurred. Eighteen attributes describe the accident as how sever the accident was, the collision type, the number of vehicles involved, the number of people involved, the number of injuries and fatalities, and if the accident involved

pedestrian and/or bicycles. Four attributes informed driver's related causes such as distracted driving (inattention), driving under influence of drugs or alcohol, speeding, and whether or not the pedestrian right of way was granted. Three attributes described the weather condition, the road condition, and the light visibility when the accident occurred. The last two attributes were miscellaneous with no clear meaning (unknown).

Attributes selection:

To select the important features that will be used in the analysis, it is important to clarify the question that needs to be answered. The question is that ***can driver related and/or the environment's related causes of car accidents predict the severity of the accident that the driver could be involved into?*** Based on the previous question, it is clear that the target feature is the severity of the accident that should be predicted i.e (***SEVERITYCODE***). This target features is already coded in the data as "**One**" for the accidents resulting in property damage with no injury or "**two**" for the accidents resulting in injuries. Table 2 shows the selected independent features that will be used to predict the target feature. Four of these features are environment's related or ER and the other four are driver's related or DR. Based on the feature selection, all the other features were dropped from the data.

Table 2: The selected independent variables to predict the target variable.

Variable	Description
Target	
SEVERITYCODE	A code that corresponds to the severity of the collision 1- prop damage 2- injury
Independent	
Environment's related or ER	
ADDRTYPE	Collision address type: • Alley • Block • Intersection
WEATHER	A description of the weather conditions during the time of the collision.
ROADCOND	The condition of the road during the collision.
LIGHTCOND	The light conditions during the collision.
Driver's related or DR	
INATTENTIONIND	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.

SPEEDING	Whether or not speeding was a factor in the collision. (Y/N)
PEDROWNOTGRNT	Whether or not the pedestrian right of way was not granted. (Y/N)

Data cleaning

Multiple processes were done to the selected feature.

- **Removing the empty or non-meaningful data:** This was done by dropping all the accidents of empty data (i.e Nan) or without clear description such as values of “Unknown” or “Other”.
- **Categorical to numerical values transformation**: Since all the values of the independent variables are of object type, the second process is to numerically coding all the text description as shown in table 3. The type of data was transformed from object to float

Table 3: The numerical coding of the data points of the independent variables

Independent variable	Values	Code
ADDRTYPE	Alley	**1.0**
	Block	**2.0**
	Intersection	**3.0**
WEATHER	Clear	**1.0**
	Partly Cloudy	**2.0**
	Overcast	**3.0**
	Blowing Sand/Dirt	**4.0**
	Severe Crosswind	**5.0**
	Fog/Smog/Smoke	**6.0**
	Raining	**7.0**
	Snowing	**8.0**
	Sleet/Hail/Freezing Rain	**9.0**
ROADCOND	Dry	**1.0**
	Sand/Mud/Dirt	**2.0**
	Wet	**3.0**
	Water	**4.0**
	Snow/Slush	**5.0**
	Ice	**6.0**
	Oil	**7.0**
LIGHTCOND	Daylight	**1.0**

	Dawn	**2.0**
	Dusk	**3.0**
	Dark - Street Lights On	**4.0**
	Dark - Street Lights Off	**5.0**
	Dark - No Street Lights	**6.0**
	Dark - Unknown Lighting	**7.0**
INATTENTIONIND	No	**0.0**
	Yes	**1.0**
UNDERINFL	No	**0.0**
	Yes	**1.0**
SPEEDING	No	**0.0**
	Yes	**1.0**
PEDROWNOTGRNT	No	**0.0**
	Yes	**1.0**

N.B. in the **UNDERINFL** variable, there were 4 types of values (“Y,N,1,0), so all “Y” values were replaced to 1.0 and all “N” values were replaced to 0.0.

- **Balancing the unbalanced data:** By running the histogram for the target variable to examine the overall distribution of values of the target variable across the data, it was obvious that the values of property damage with no injury (1) were almost double of the values of injury (2) as shown in figure 1. It is recommended to balance the data to avoid any biasness. To do that, the observations “1” in target variable were randomly shuffled and under sampled by randomly deleting some of the observations from that majority class, so the observations of the majority class can match the numbers with the minority class. Figure 2 shows the histogram of the target variable observations after balancing.

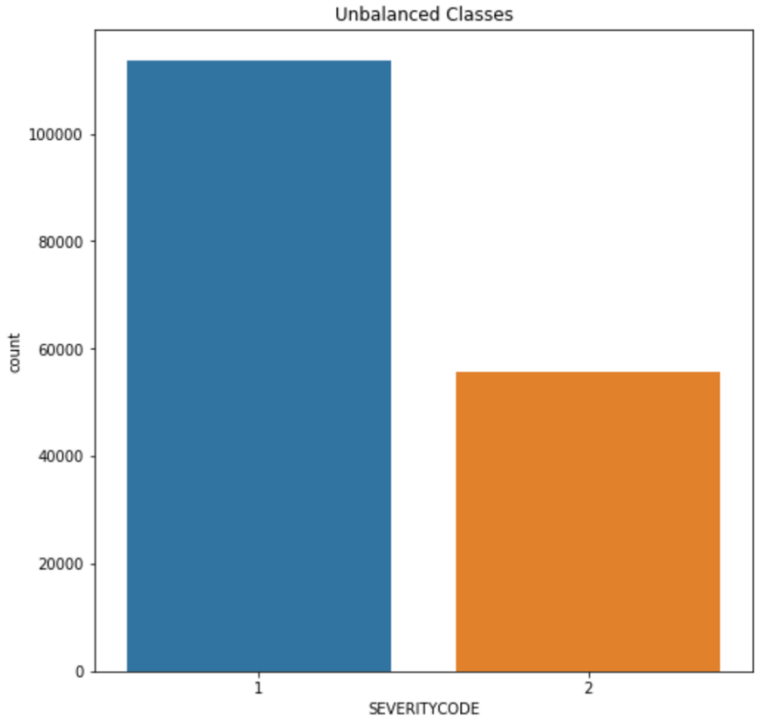

Figure 1: The histogram of the unbalanced distribution of the target variable observations (The severity of collision); 1: property damage with no injury, 2: injury.

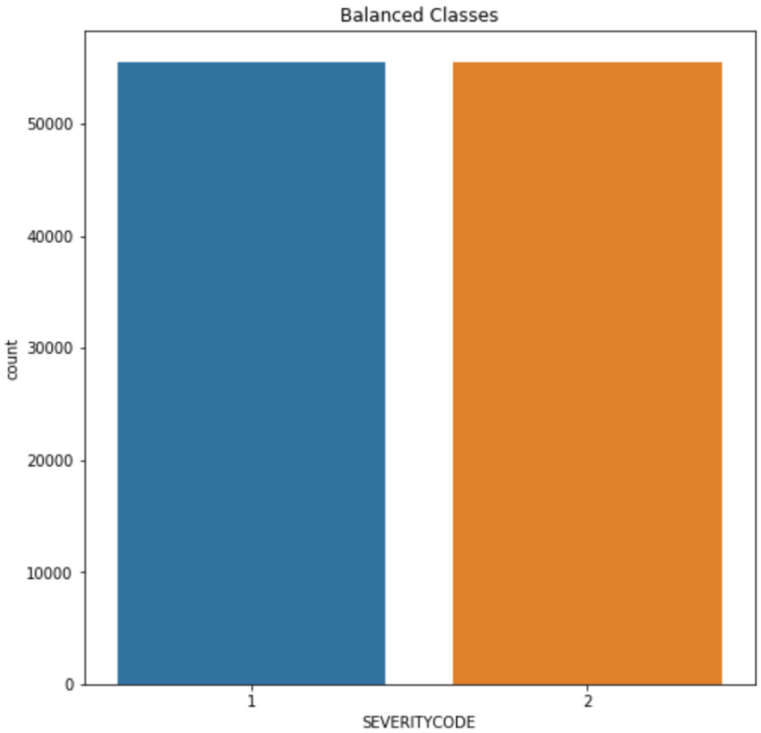

Figure 2: The histogram of the balanced distribution of the target variable observations (The severity of collision); 1: property damage with no injury, 2: injury.

Building a model

Four classification models were examined for the best accuracy to predict the target observation or the severity of the car accident after being trained using the training data. The four models are;

- K nearest neighbors (KNN)
- Support vector machine (SVM)
- Logistic regression (LR)
- Decision trees (DT)

To know the best features that will be accurately used to predict the severity of car collision, the 8 selected features were portioned into three feature sets;

- Environment related features (X_ER) (ADDRTYPE, WEATHER, ROADCOND, LIGHTCOND).
- Driver related features (X_DR) (INATTENTIONIND, UNDERINFL, SPEEDING, PEDROWNOTGRNT).
- All features combined (X_all).

Each feature set was examined through all classification models to get the best model and feature set of the highest accuracy of prediction. The process started by assigning the features into independent variable X based on the corresponding set as (X_ER, X_DR, and X_all), and dependent variable y. Then sklearn/preprocessing library was used to normalize the dependent variable X. For each X_ER, X_DR, and X_all, 80% of the data was used as training sets (X_train and y_train), and the remaining 20% of the data was assigned as test sets (X_test and y_test) to evaluate the model after being trained using the training sets.

Results

KNN model: Before running the model, the K of the best accuracy was calculated for all the sets of features. The best K was 8, 1, and 12 for X_ER, X_DR, and X_all, respectively, as shown in figure 3. Using the best K, the KNN model was conducted using the training set of data, and the model was evaluated using the test data set. The accuracy of the model for training and test data sets was calculated as well as the overall accuracy as indicated by F1-score as shown in table 4. Using all features to train the model (X_all), the KNN model showed the highest accuracy indicated by F1-score (0.583) compared to the other sets of features (X_ER and X_DR). While the lowest accuracy of the KNN model was reported when the driver related features (X_DR) were used to train the model, the environment related features (X_ER) showed F1-score that was very close to the highest F1-score reported by the all features, which implicated a stronger influence of the X_ER to predict the severity of the car collision compared to X_DR.

A B C

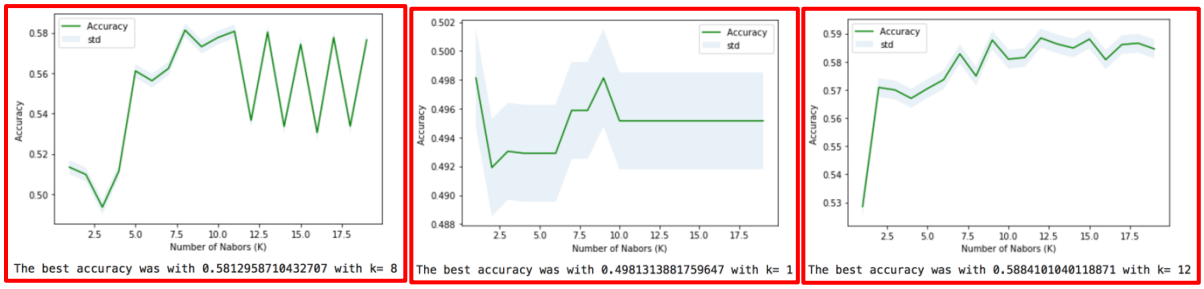

Figure 4: The best K giving the highest accuracy of the KNN model for A) X_ER, B) X_DR, and C) X_all.

Table 4: The accuracy of the KNN model across all the sets of features

Features set	Accuracy for training set	Accuracy for test set	F1-score
X_ER	0.57	0.58	0.57
X_DR	0.50	0.49	0.33
X_all	0.58	0.58	0.58

SVM model:

The accuracy of the SVM across the different feature sets was comparable to what was obtained by the KNN. The F1 scores were 0.5896, 0.4542, and 0.6030 for X_ER, X_DR, and X_all, respectively. Again the SVM model showed the same trend of a higher accuracy when all features were used and a lower accuracy when the driver related features (X_DR) were used. The confusion matrix were implemented to visually show accuracy indicated by increasing the number of the true target observations of the severity of the car collision. This was obvious for X_all as shown in figure 5.

LR model:

As shown in table 5, the same trend of higher accuracy indicated by a higher F1-score (0.603) was found when all features were used to train the LR model. Moreover, the same set of features (X_all) showed also a lower log loss (0.653). The confusion matrices for the LR model for all the sets of features were shown in figure 6.

Table 5: The accuracy of the LR model across different sets of features

Feature set	F1-score	Log loss
X_ER	0.5899	0.673
X_DR	0.5333	0.6705
X_all	0.6026	0.6534

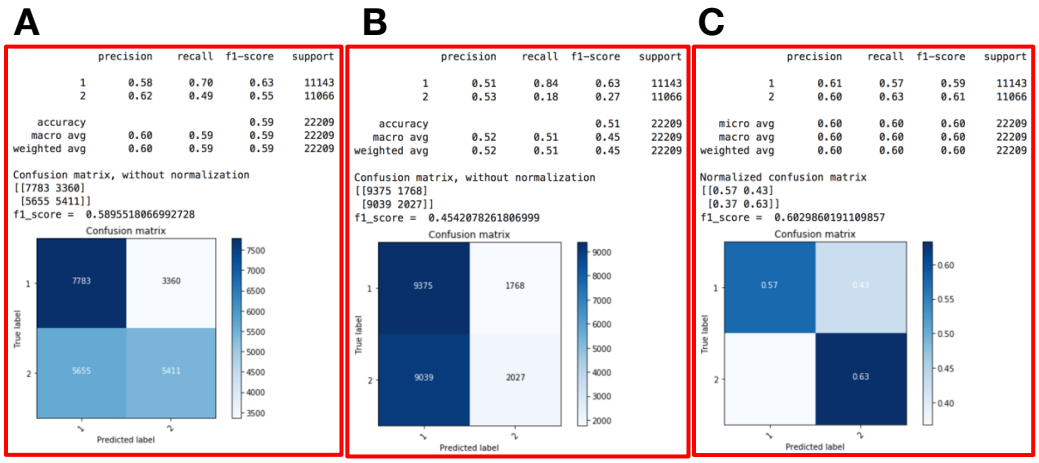

Figure 5: The confusion matrix for the SVM model showing the precision and F1- scores for A) X_ER, B) X_DR, and C) X_all.

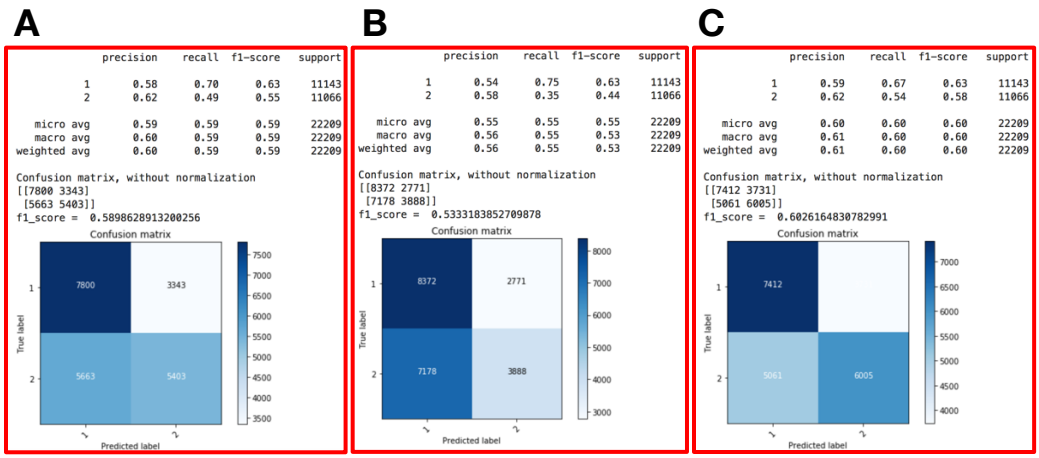

Figure 6: The confusion matrix for the LR model showing the precision and F1- scores for A) X_ER, B) X_DR, and C) X_all.

DT model:

The DT model showed also the same trend of higher accuracy (0.60) when all features were used compared to a lower accuracy (0.55) when the driver related features were used. However, the accuracy of the DT model for the environment related features (X_ER) was close to the highest accuracy (0.59). The construction of the DT (Figure 7) of all features showed that light condition (LIGHTCOND), road condition (ROADCOND), driving under the influence of drugs or alcohol (UNDERINFL), and not giving the right for pedestrian (PEDROWNOTGRNT) were the only features that were able to get the entropy to zero and getting some pure nodes as shown in orange highlight in figure 7.

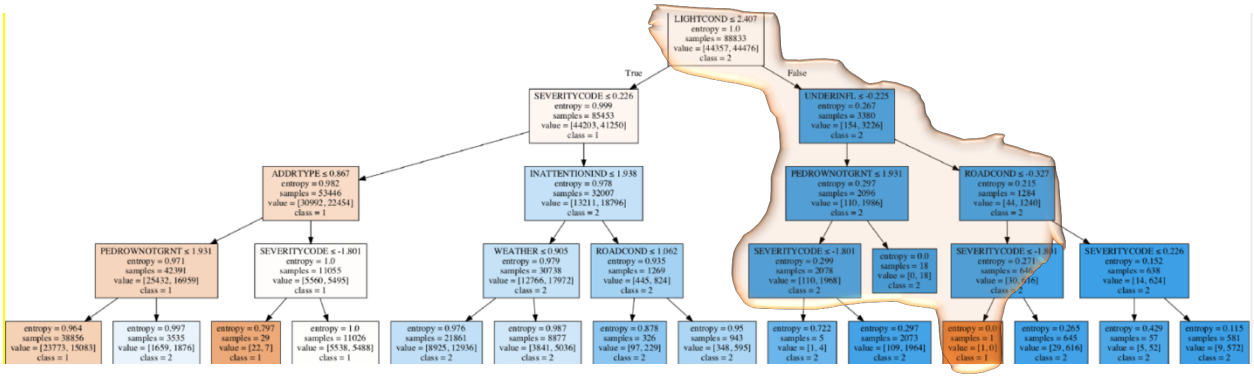

Figure 7: The decision tree (DT) of all features used to train the model.

Discussion and conclusion:

By comparing the accuracies of all models across the different sets of features, it is clear that SVM, LR, and DT were the best models having higher F1-scores (0.60). This highest accuracy was obtained by using the environment and driver related features combined (X_all). However, the F1-score is not that high for the classifier to have to be able to predict the severity of the car collision. This notion was supported by the DT model, where it was difficult to get pure nodes for target observations (1 vs 2 or property damage vs injury) of the severity of the car collision. One of the problem is the absence of the clear separation between the observations of the target variable. The collision resulting in injury will have also a property damage. In other words, all the injury class observation are of property damage type, which could explain why the raw data were unbalanced and biased toward the property damage class. The classes of the target variable should be balanced and exclusively separated. Indeed, one step of the data cleaning was balancing the data, but this was done by under sampling the majority class of the target variable, which resulted in losing a lot of data points. Also, the under sampling of the majority class of the target variable was done by randomly removing some data points which could be important for the efficiency of the classifier. In conclusion, the environment features such as the road condition, weather, and the degree of visibility are very important features to predict the severity of the car

collision. Sever weather and road conditions as well as low visibility could escalate the car accident. The driver's behavior seems to have a lower impact to predict the severity of the car collision, but it certainly strengthened the accuracy of the classifiers used to predict the severity of the car collision when it was combined with the environment features.

Table 7: The accuracy of all models across different sets of features:

Feature set	F1-score (KNN)	F1-score (SVM)	F1-score (LR)	Log loss (LR)	Accuracy of DT
X_ER	0.57	0.59	0.59	0.67	0.59
X_DR	0.33	0.45	0.53	0.67	0.55
X_all	0.58	0.60	0.60	0.65	0.60