Predicting the Severity of Road Accidents

Shilpa Banerjee

September 23,2020

1. Introduction

1.1. Background

The Seattle community has expressed concern over the serious road accidents taking place in certain road junctions. Incidents of vehicular collisions as well as vehicles hitting pedestrians or cyclists have been reported.

The increasing number of serious accidents will eventually bring down the road safety score of this suburb there being leading to unpleasant consequences. Traffic jams caused by accidents are a nuisance for all those on the road.

1.2. Problem

Say you are driving to another city for work or to visit some friends. It is rainy and windy, and on the way, you come across a terrible traffic jam on the other side of the highway. Long lines of cars barely moving. As you keep driving, police car start appearing from afar shutting down the highway. Oh, it is an accident and there's a helicopter transporting the ones involved in the crash to the nearest hospital. They must be in critical condition for all of this to be happening. Now, wouldn't it be great if there is something in place that could warn you, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to.

1.3. Interest

It would be great if there is something in place that could warn people, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be.

2. Data acquisition and cleaning

2.1. Data sources

I have used the dataset provided in the example. This is the Seattle accidents database which discusses the fatality of accidents based on a set of conditions. The fatality is measured using attribute SEVERITYCODE has been given values 1 or 2.

SEVERITYCODE	**Description**
1	Property damage
2	Injury

This score is based on several factors some of which are mentioned below.

Location	Road Condition
Weather Condition	Junction
Car Speeding	Number of people involved
Light Conditions	Number of vehicles involved

There is a total of 37 attributes which are a mix of numerical and categorical datatypes.

2.2. Data Cleaning

This data ranges from years 2004 to 2020 so it can be considered up to date for developing the model. There are however quite a number records with null values.
Out of 37 columns we can see redundant columns as we have categorical codes and their respective descriptions (SEVERITYDESC, SDOT_COLDESC, ST_COLDESC, LOCATION). These description columns have been removed.
There are a few ID columns (OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS, INTKEY, SDOTCOLNUM, ST_COLCODE, SEGLANEKEY, CROSSWALKKEY) as well which are not useful for our purpose and removed from the dataset.
The columns (INATTENTIONIND, UNDERINFL, EXCEPTRSNCODE, EXCEPTRSNDESC, PEDROWNOTGRNT, SPEEDING) has totally or mostly null values hence they are removed.

Some columns (INCDTTM, JUNCTIONTYPE) have redundant information and hence have been excluded.
Original dataset comprised of 37 columns and 194673 records. After cleaning up we have 13 columns and 184146 records remaining.

2.3.Feature Selection

The column SEVERITYCODE is our y value which has to be predicted.

The columns WEATHER, ROADCOND and LIGHTCOND are closely correlated. Hence, I have considered only one out of these three for the model.

HITPARKEDCAR is correlated to COLLISIONTYPE as we are indicating the same in that as well.

Final Feature selection is as follows:

Column	**Used in Model**	**Description**	**Reason for Exclusion**
ADDRTYPE	Included	This describes the junction where accident occurred (Alley,Block,Intersection)	
COLLISIONTYPE	Included	This describes the collision (Angle,Head on, Side swipe, Rear end, Parked car etc)	
PERSONCOUNT	Included	The total number of people involved in the collision	
VEHCOUNT	Included	The number of vehicles involved in the collision.	
ROADCOND	Included	A description of the road conditions during the time of the collision.	
PEDCOUNT	*Excluded*	*The number of pedestrians involved in the collision*	*Correlated to COLLISIONTYPE*
PEDCYCLCOUNT	*Excluded*	*The number of bicycles involved in the collision.*	*Correlated to COLLISIONTYPE*
HITPARKEDCAR	*Excluded*	*Whether or not the collision involved hitting a parked car*	*Correlated to COLLISIONTYPE*
WEATHER	*Excluded*	*The condition of the weather during the collision.*	*Correlated to WEATHER*
LIGHTCOND	*Excluded*	*The light conditions during the collision.*	*Correlated to WEATHER*
X	*Excluded*	*X – coordinate of the collision location*	*Not useful for model, may use for map*
Y	*Excluded*	*X – coordinate of the collision location*	*Not useful for model, may use for map*
INCDATE	*Excluded*	*Date of Incident*	*Not useful for model, used for studying spread of data.*

3. Exploratory Data Analysis

3.1. Calculation of target variable

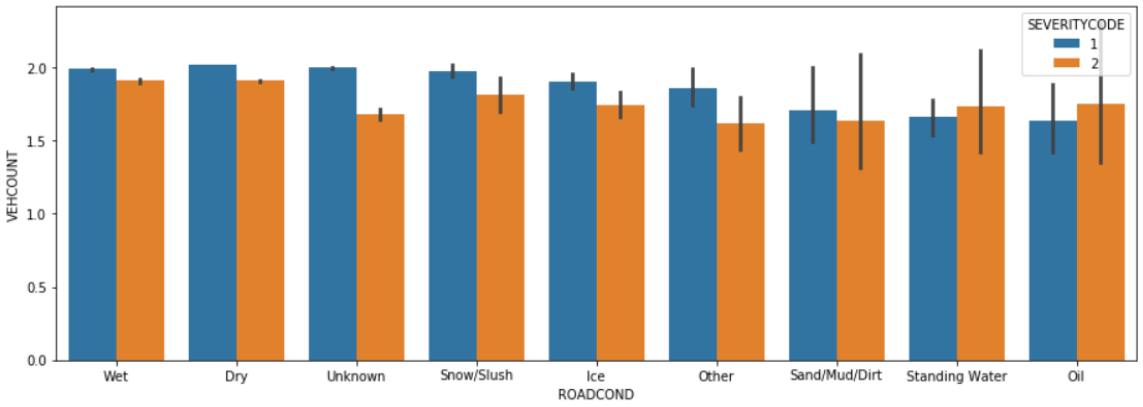
Severity of the accident is the target variable that needs to be predicted. It has values 1 or 2 explained below:

SEVERITYCODE	Description
1	Property damage
2	Injury

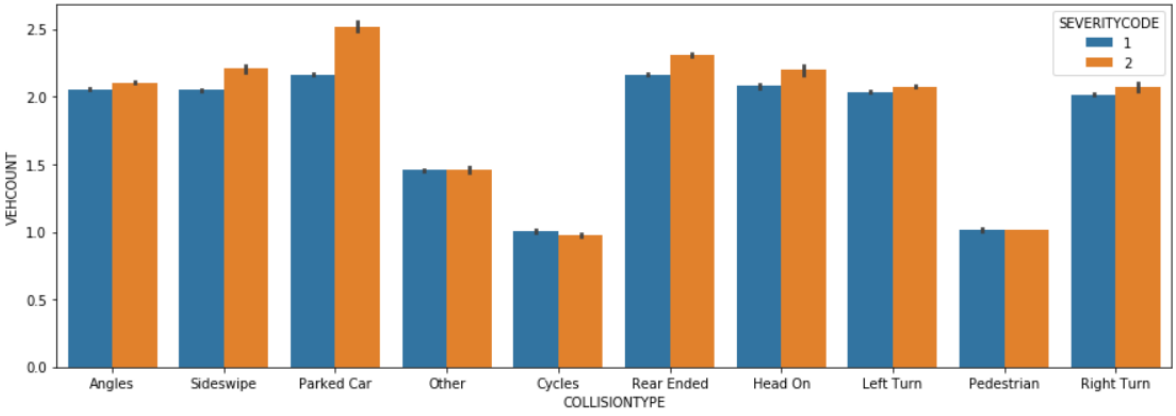
I have analysed the severity of the accident under the following two groups:

3.2. Total number of vehicles involved in collision:

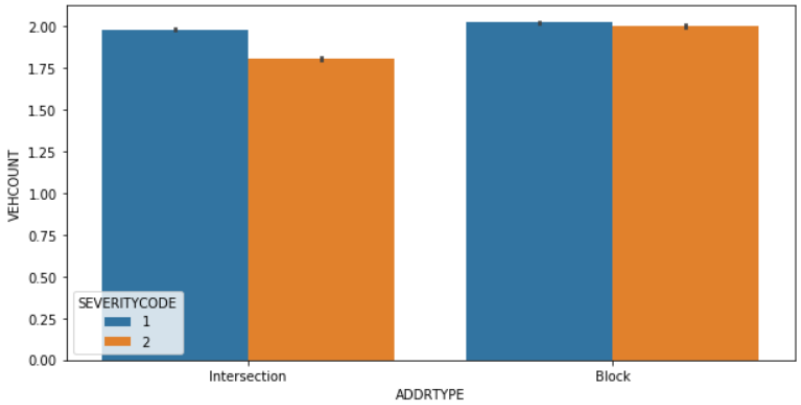
3.2.1. Relationship between accident severity, road condition and total number of vehicles involved in collision



3.2.2. Relationship between accident severity, type of accident and total number of vehicles involved in collision

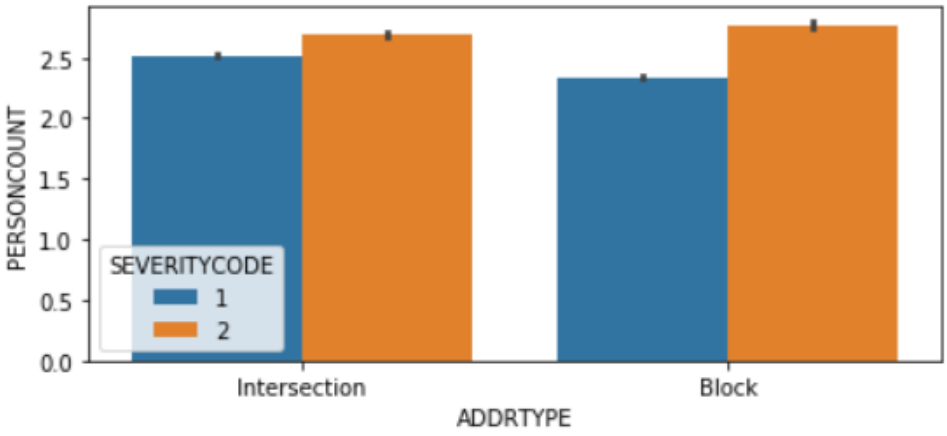


3.2.3. Relationship between accident severity, location of accident and total number of vehicles involved in collision

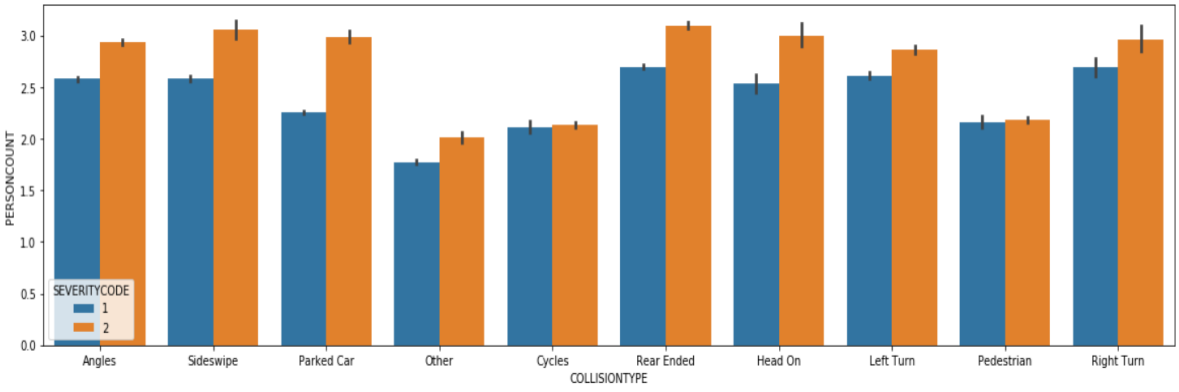


3.3.Total number of people involved in collision:

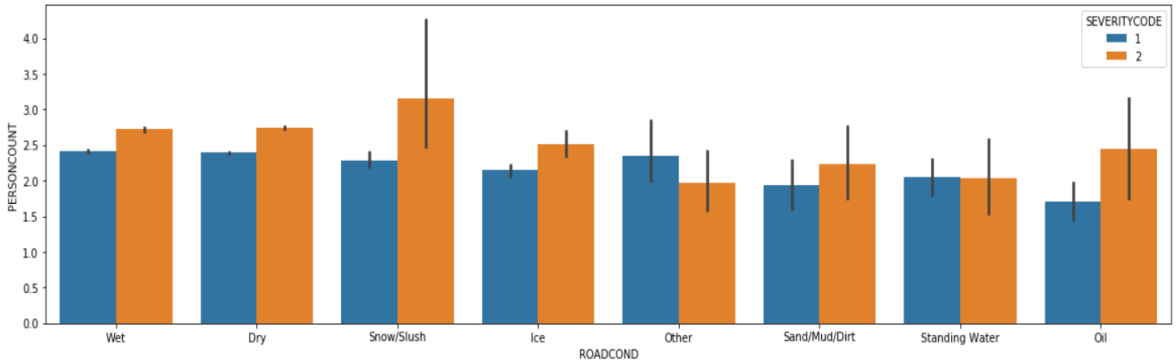
3.3.1. Relationship between accident severity, location of accident and total number of people involved in collision



3.3.2. Relationship between accident severity, type of accident and total number of people involved in collision



3.3.3. Relationship between accident severity, road condition and total number of people involved in collision



We observe an even spread of values as seen from the bar charts. Each of the features VEHCOUNT, PERSONCOUNT, ADDRTYPE, ROADCOND and COLLISIONTYPE are significant contributors to the target variable SEVERITYCODE.
In other words, SEVERITYCODE is dependent on all these factors.
Hence, we have considered all of them for building the model.

4. Predictive modelling

As we have to predict categorical variable, I am using classification algorithms.

I have built the models using the following classification algorithms and the efficiency has been compared using Jaccard similarity score and F1-score

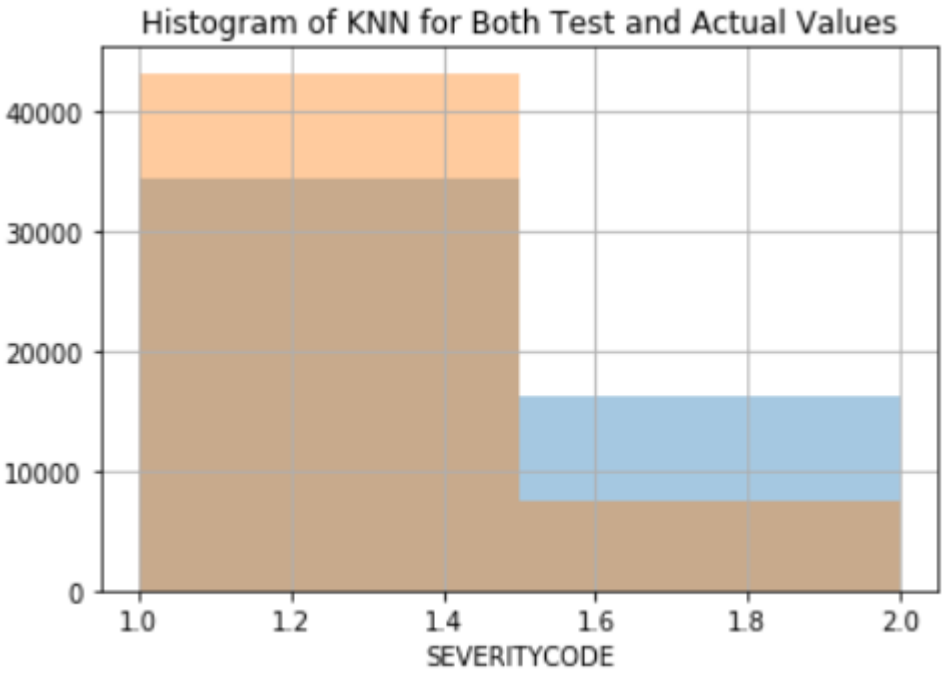
The histogram in each of the algorithms shows a comparison between actual values of SEVERITYCODE from the test set as compared to the predicted values by that particular algorithm.

4.1.K- Nearest Neighbours (KNN)

Using KNN with k = 6 we get best accuracy of 73.41%

Jaccard Similarity Score = 73.51%

F1 Score = 70.05%

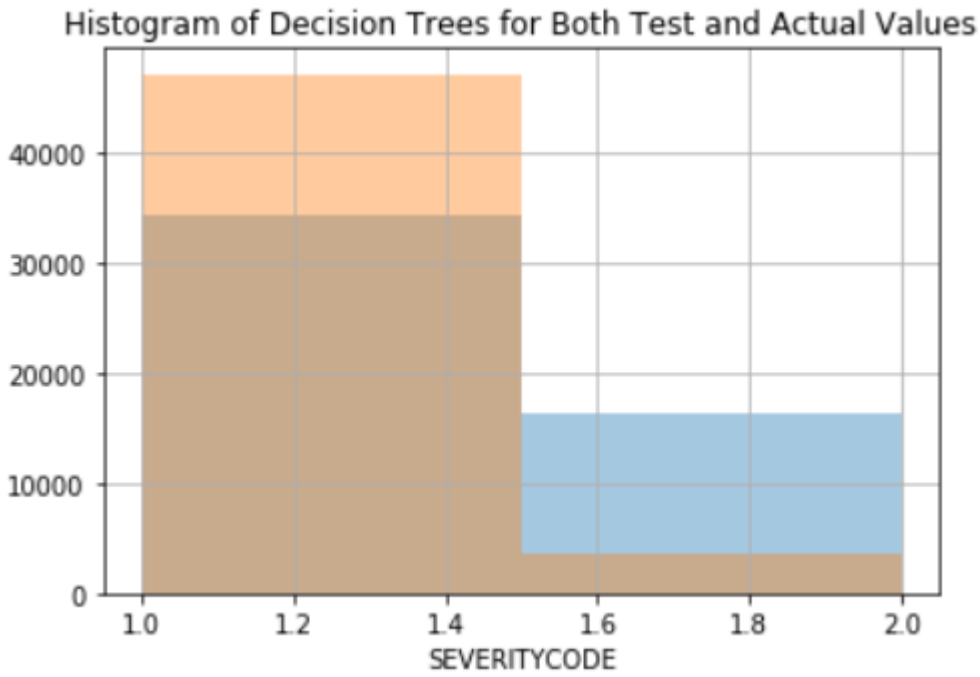


4.2.Decision Trees

Using decision trees with entropy = 4 we get best accuracy of 73.08%

Jaccard Similarity Score = 73.08%

F1 Score = 66.42%

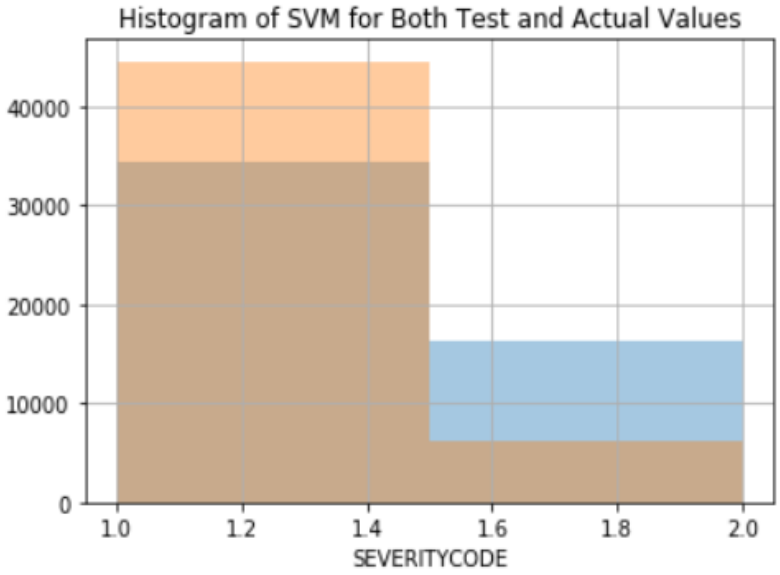


4.3.Support Vector Machines (SVM)

Using SVM we get the following accuracy:

Jaccard Similarity Score = 74.01%

F1 Score = 69.81%



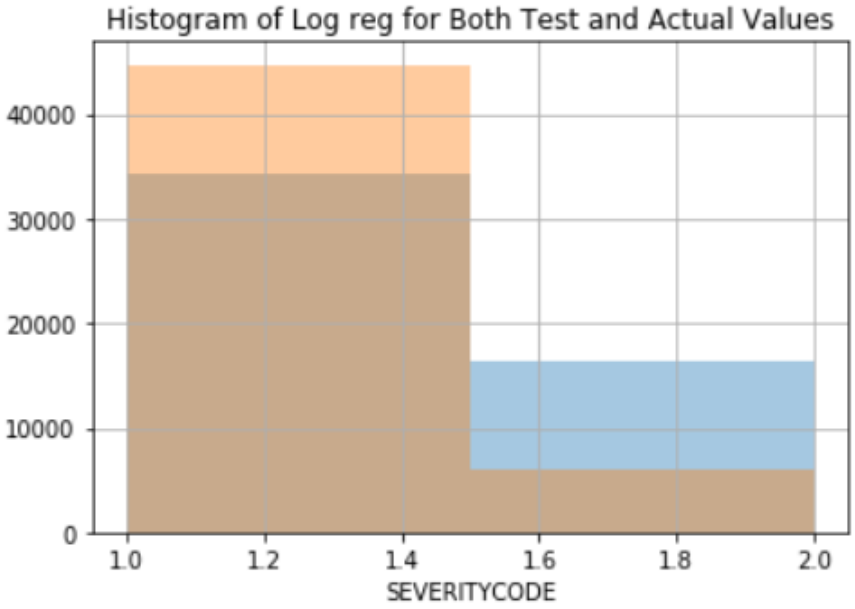
4.4.Logistic Regression

Using Logistic Regression, we get the following accuracy:

Jaccard Similarity Score = 73.81%

F1 Score = 69.43%

Log Loss = 51.59%



5. Conclusions

For the given dataset, the SVM algorithm gives the best accuracy as shown by Jaccard score of 74%.

Other algorithms are not too far behind with Jaccard scores of around 73 %

6. Inference

As SVM has the best accuracy we can use it to make predictions of accident severity.

Given a certain location, conditions of the road, type of collision that occurred , number of vehicles and people involved in the accident we can predict if it will lead to just property damage or injuries as well.