Predicting the Severity of Traffic Accident

Bowen Deng

August,2020

Contents

Abstract

This project is a machine learning modeling and data analysis using the Seattle vehicle crash dataset provided by Kaggle.

1. Introduction

1.1 Background

Transportation is a very important part of our lives. People have many transportation options available to them, and usually many people also choose to drive their own cars. Maybe it's a sunny weekend and you want to visit friends and family; or maybe it's a workday business trip to somewhere.and on the way, you encounter a terrible traffic jam where the long queue of cars on the other side of the highway can barely move. As you continue driving, police cars begin to appear from the distance, shutting down the highway. Oh, it's an accident, and a helicopter is transporting the people involved in the crash to the nearest hospital. They must have been in critical condition for all this to happen.Traffic accidents are hard to predict, but they are endless.

1.2 Problem

Now, what if something could warn you, given the weather and road conditions, of the possibility of getting into a car accident and how serious it could be, so that you could drive more carefully and even change your trip. Wouldn't it be great if the above predictions could be made. You would be able to keep yourself safe and avoid experiencing traffic jams. So this is the problem that this subject is trying to solve.

1.3 Target Audience

The intended audience for this project will be the Department of Transportation ,Traffic Radio, Navigation Software.Due to the danger of vehicle collisions, providing solutions that may reduce the amount of accidents can significantly improve the quality of life of pedestrians & overall ensure public safety.

1.4 Data

Now here's a huge data source obtained that I believe we can use to construct supervised learning models in machine learning, such as using features: location,Road Condition,Weather Condition,Function,Car Speeding,Number of people The data is then used to match realistic predictions of whether and what kind of traffic accidents are likely to occur. Most importantly, it contains a severity code that ranges from 0 (unknown) to 3 (fatality). Being able to use the various features within the dataset to better predict this the level of severity of the collision.

2. Methodology

2.1 Data Source

The dataset for this topic is derived from Kaggle and contains Seattle vehicle crash records for the intervening years 2004 to 2020.

2.2 Data Cleaning

First, after importing this CSV file we can see that the dataset has 38 columns and over 200,000 rows. We use the command to query the dataset and see that it is incomplete and has missing values, but also that there are problems with the formatting of some of the columns, so we first clean the data.
For example, in the 'SEVERITYCODE' column we can see that there are 5 collision types present, however there is a value of '2b' which is different from the others. To make the calculation easier, we replace it with '2.5' so that it can be of type float with any other value.

Then, we find that there are two columns for time, both with the same meaning. The time can also be our eigenvalue to predict whether the collision result is related to the day of the week. So first standardize it with 'to_datatime' and then use 'dayofweek' to convert it to be specific to the week of the week. The number type is easy to calculate, which is the purpose of the conversion. It is then added to one and recorded as one to seven.

Next came the most important part, the treatment of the missing values. After looking at this dataset, it was found that almost most of the column values are discrete rather than continuous, so the treatment of averages, maxima, etc. is not well done. And a query of the columns with missing values found that most of the column types are more average, so the most frequent values are also difficult to assess. Since there are more than 200,000 rows in the dataset, after discarding the rows containing missing values, it is found that there are 180,000 rows in the dataset, so such processing can be used.

2.3 Feature Selection

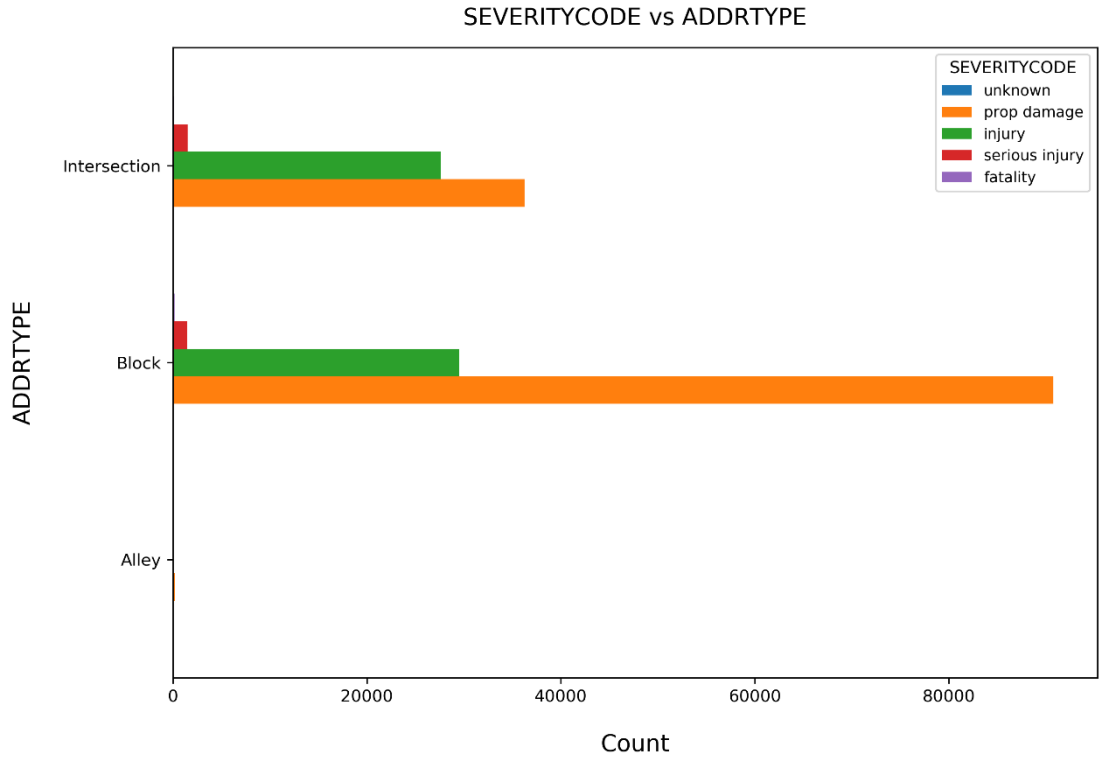
Now, let's move on to the feature selection phase. Looking at all the columns, we see that not all of them are suitable to be selected as features for modeling. Let's recall that our goal is to use this dataset to predict whether a traffic accident will occur and what the extent of the collision will be if it does. The audience is traffic data, navigation software, etc., so that whether a driver has been drinking cannot be predicted in advance. Also, the number of people involved in collisions, injuries, and fatalities are ex-post data and cannot be characterized as ex-ante predictions.

So, in summary, we can use e.g. weather (weather is a good feature, e.g. rain will definitely affect the driver's vision), road conditions (road conditions are definitely related to the collision), lighting conditions (visibility is also a factor that affects driving), speeding (this definitely needs to be selected, and navigation software can also alert the driver based on speed), intersection category (navigation software can also). Based on the intersection category to alert drivers to which locations are high accident locations), and the date (which has been addressed above and converted to a day of the week, which lets us know which days of the week are high accident days and may be peak travel times). Finally, select our target: "SEVERITYCODE", which represents whether or not an accident occurred and the extent of the collision if it did.

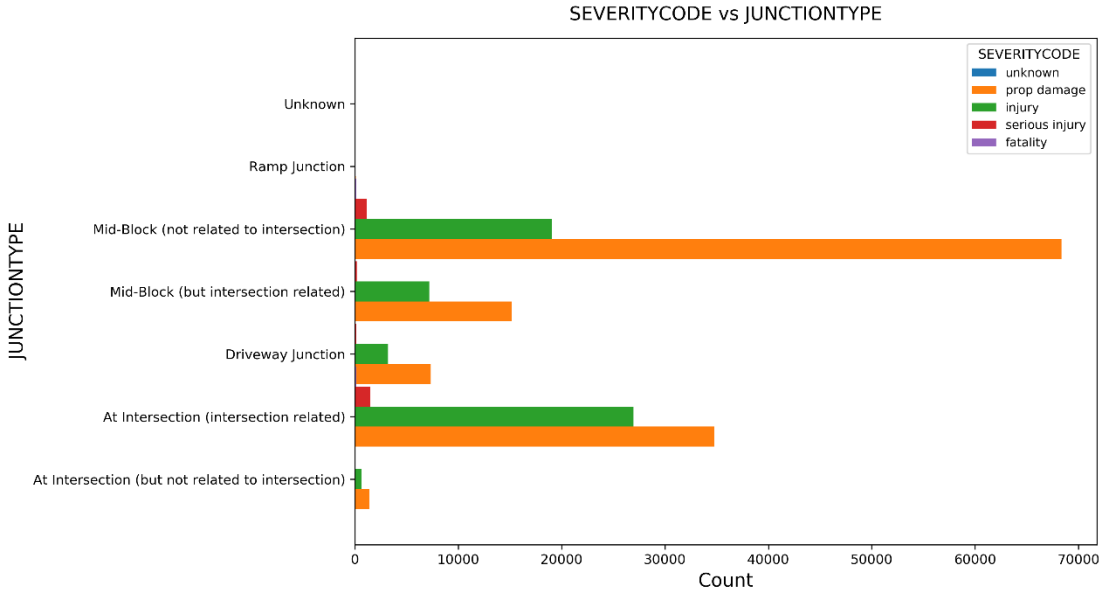
2.4 Exploratory Data Analysis

To re-validate our point, we draw bar graphs for the column 'SEVERITYCODE' and the other feature columns.

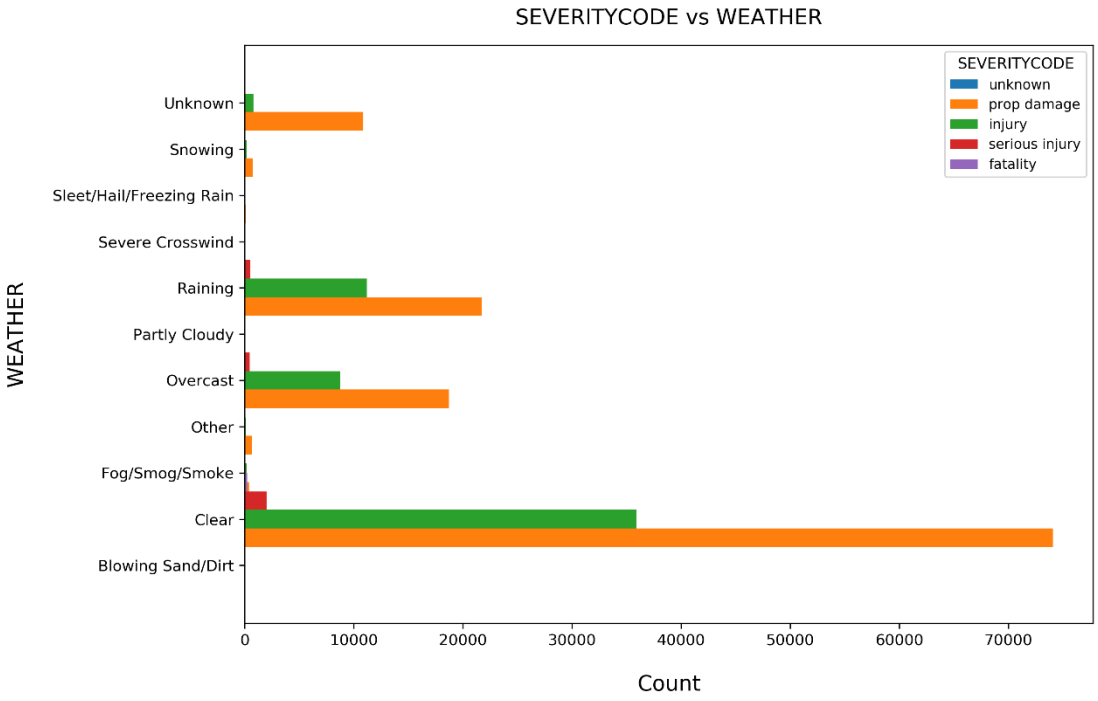
2.4.1SEVERITYCODE VS ADDRTYPE



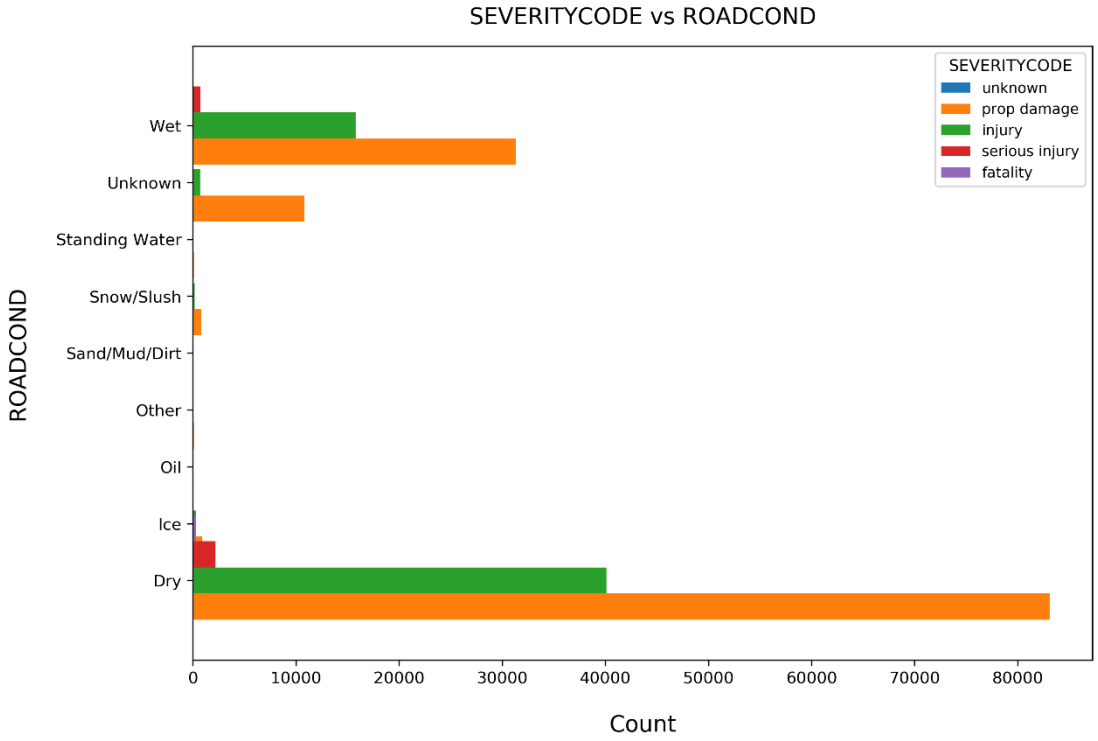
2.4.2SEVERITYCODE VS JUNCTIONTYPE



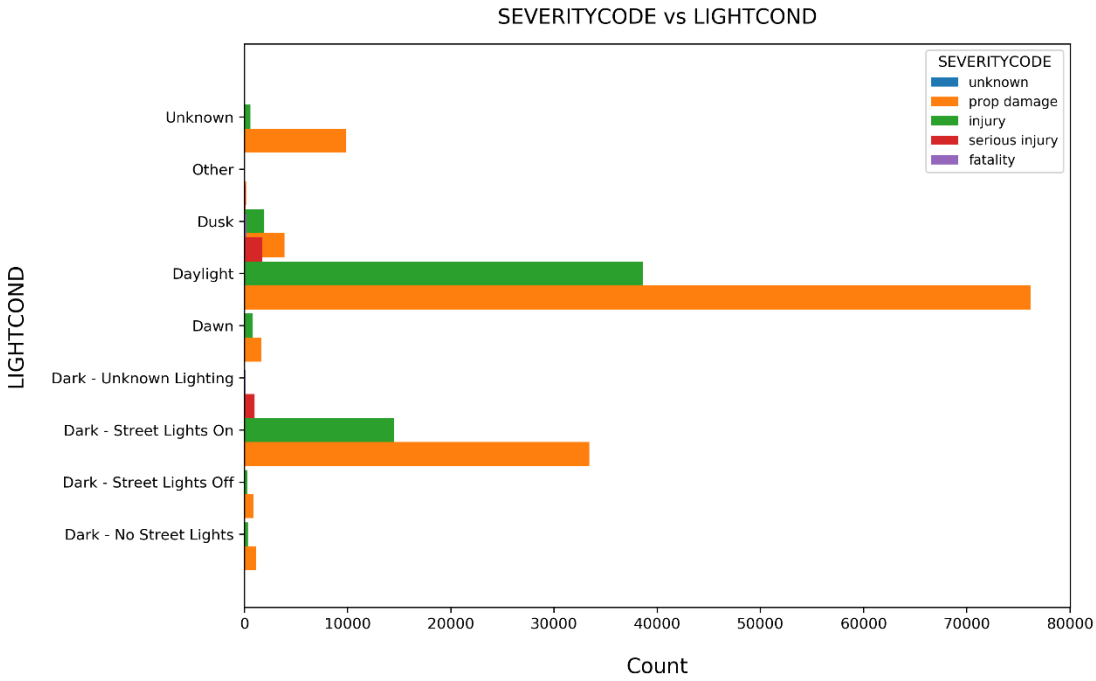
2.4.3SEVERITYCODE VS WEATHER



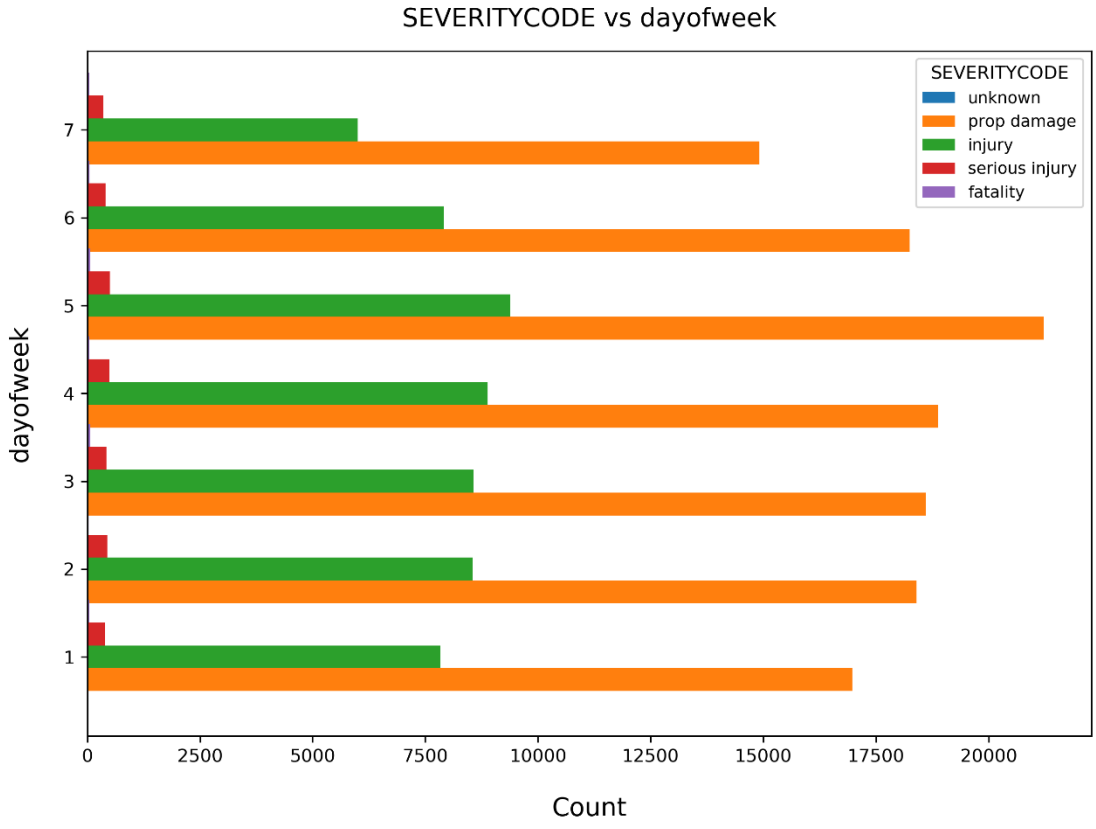
2.4.4SEVERITYCODE VS ROADCOND



2.4.5SEVERITYCODE VS LIGHTCOND



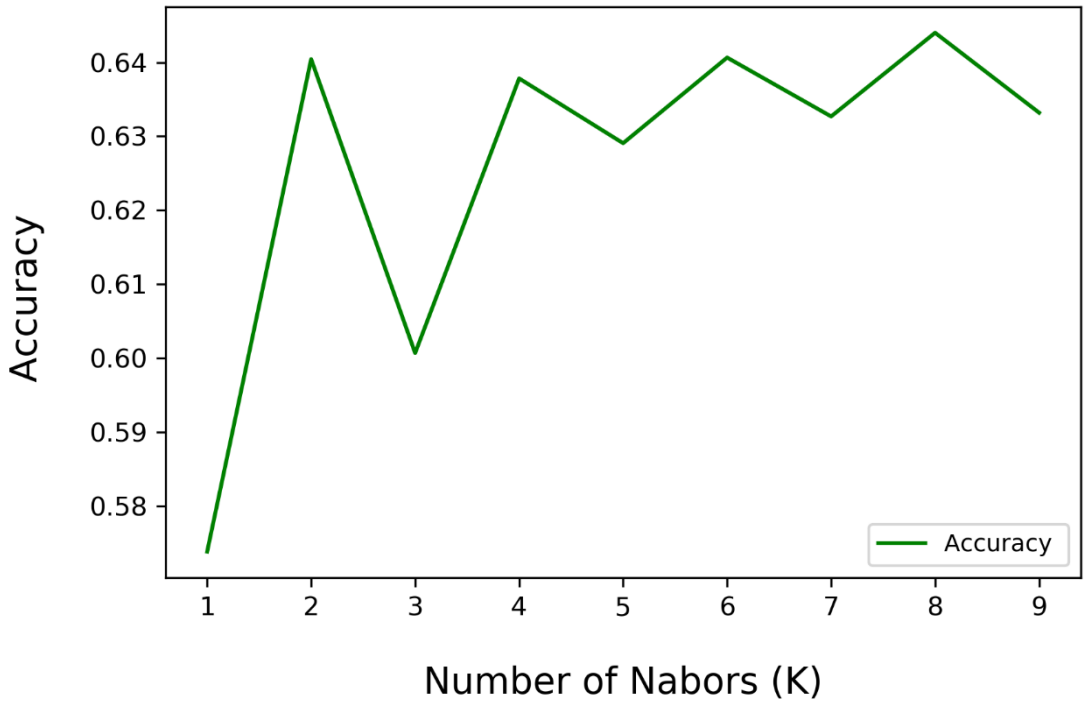
2.4.6SEVERITYCODE VS dayofweek



2.5 Modeling

Since this is a classification problem, I use four models - KNN, decision tree, SVM, and logistic regression. The column values of the raw data are almost always character-based, so each type is assigned a specific value as a way to transform the type and facilitate modeling calculations. After splitting the dataset into a training set and a test set, modeling with the training set and testing with the test set, the Jaccard coefficients for each of the four models were derived as follows.

2.5.1KNN



Jaccard= 0.6440023436667732
F1_score= 0.587870335832594

2.5.2Decision Tree

Jaccard= 0.6765207201448812
F1_score= 0.5459882234510015

2.5.3SVM

Jaccard= 0.6765207201448812
F1_score= 0.5462335783770929

2.5.4Logistic Regression

Jaccard= 0.6732449131777991
F1_score= 0.5484759130295178

3. Results

In summary, the models of decision trees and SVMs fit better relative to the models of KNN and logistic regression.

4. Discussion

The dataset has a number of limitations, such as a number of missing values and problems with pre-classification of values. More data would help to significantly improve the performance of the model.

5 Conclusions

In this study, I analyzed the relationship between Seattle vehicle crash chances and crash levels and crash day data. I found that the type of intersection, date, weather on the day, road conditions, and light were the most important characteristics of the collision. I developed a classification model to predict whether drivers would collide and the extent of the collision. This model is useful in a variety of ways, including big traffic data and navigation software. For example, it can be used to remind drivers to drive safely or to change their travel dates.