Predicting the Seriousness of Car Accidents in Seattle, WA

Tony Lizza

September 15, 2020

1. Introduction

1.1. Background

Driving is the most dangerous activity that most people engage in on a regular basis. Traffic accidents remain one of the largest causes of death and serious injury for Americans. In addition to the significant human cost, motorists and logistics companies alike face costly delays due to the impact of these serious crashes on traffic. Lane closures may be required for long periods of time to deal with the results of the accident, for example, to transport injured and/or clear accident wreckage.

1.2. Problem

The business value we are attempting to provide is to answer the question "What is the best way to predict (and therefore avoid) serious car accidents when driving?" The goal of this exercise is to predict the severity of a given road collision when provided with certain details regarding the accident (e.g., weather and the road conditions) to allow drivers to drive more carefully or modify their travel plans in response to avoid encountering accidents.

1.3. Interest

This topic will be of interest to motorists and transportation/logistics companies alike who have an interest in avoiding circumstances that lead to serious car accidents, both to avoid being a part of such an accident, as well as to avoid the resulting serious traffic delays arising from the accident.

2. Data Acquisition and Cleaning

2.1. Data Source and Description

The data used for the analysis was supplied by the Seattle Department of Transportation (SDOT) and consists of tabular data of 194,673 road collisions and their details from the Seattle, WA area collected between 01-01-2004 and 05-20-2020. The data includes location (lat, long, and address), type of collision, number involved (for persons, pedestrians, cyclists, and vehicles), incident date, whether the driver was inattentive, whether the driver was under the influence, and whether the driver was speeding, road conditions, weather conditions, and time of day. The dependent variable is a Boolean value indicating whether the accident was a serious one.

2.2. Data Cleaning

The data was mostly well-formatted, however some additional cleaning and processing was required. Several of the examples were missing values for certain columns in the dataset. For

example, there were several thousand missing values for the ‘under the influence’ indicator. In cases where the column was a binary value, I replaced the null value with the most common value for the column. In cases where the column values were categorical, I converted blank or null values to ‘Other’ or ‘Unknown’ as a first step for later processing (i.e., replacement with one-hot vector encoding). I also standardized the values of binary variables from Y/N or other values to 0/1. The target variable was present for all samples, and so removal of individual samples was not required. Finally, a small number of duplicate samples were found in the data, which were removed.

2.3. Feature Selection

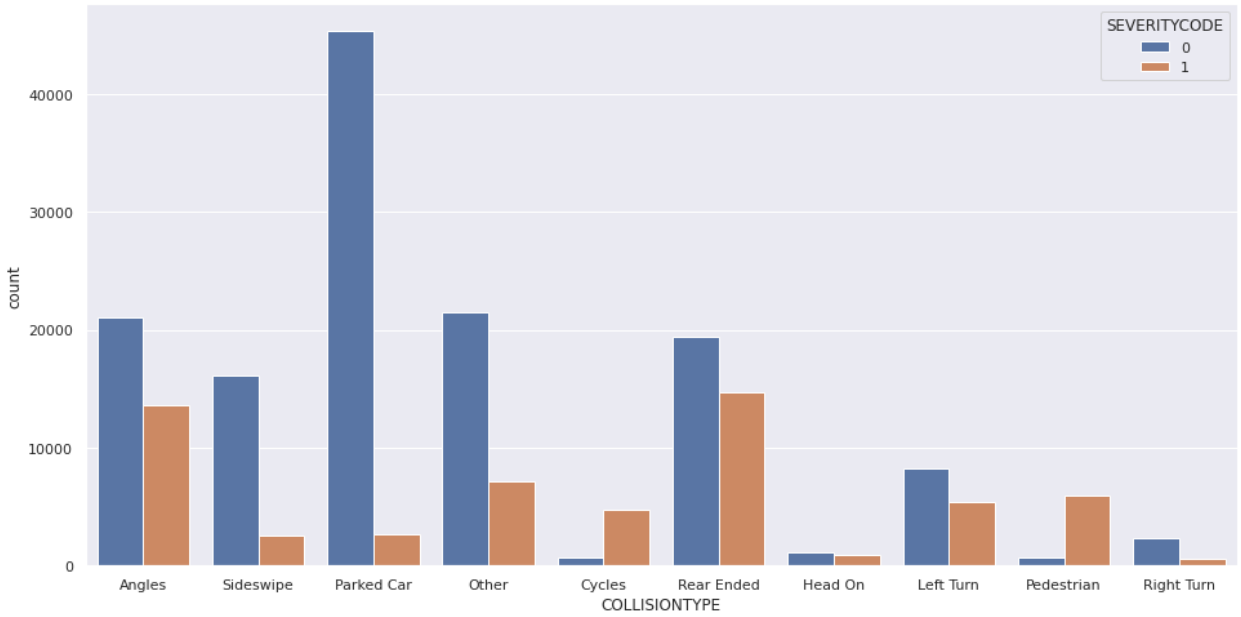
After cleaning, the data consisted of 194,621 samples and 37 features in the dataset. On examining each of the features, it was clear that there were database-specific unique identifiers, as well as redundant features, both of which were not necessary or desirable for creation of the model. OBJECTID, INCKEY, COLDETKEY, REPORTNO, and other fields are unique identifiers that appear to be significant either for the database or for law enforcement and won’t be useful to our analysis. SEVERITYCODE.1 is a duplicate and SEVERITYDESC is a redundant field that describes SEVERITYCODE. From the given data, I initially selected the following fields to create the model:

- ADDRTYPE: Type of location where the crash took place, e.g., intersection.
- COLLISIONTYPE: Type of collision, e.g., head-on.
- PERSONCOUNT: No. of persons involved
- PEDCOUNT: No. of pedestrians involved
- PEDCYLCOUNT: No. of cyclists involved
- VEHCOUNT: No. of vehicles involved
- SDOT_COLDESC: Description of accident. Note: This is a list of values.
- INATTENTIONIND: Whether or not the driver was inattentive at the time of the crash
- UNDERINFL: Whether or not the driver was under the influence at the time of the crash
- WEATHER: Weather at the time of the crash
- ROADCOND: Road conditions at the time of the crash
- LIGHTCOND: Lighting conditions at the time of the crash, e.g., whether daytime, nighttime with streetlights, etc.
- PEDROWNOTGRNT: If a pedestrian was involved in the crash, whether or not they had the right of way.
- SPEEDING: Whether or not speed was a factor in the crash
- HITPARKEDCAR: Whether or not motorist hit a parked car

3. Exploratory Data Analysis

Looking at the data, we can see that we have 136,444 non-serious accidents and 58,177 serious accidents. This means that the data is imbalanced. I then reviewed the data for interesting correlations. The first data point I looked at was COLLISIONTYPE. From the graph below, it seems that many of the COLLISONTYPE values have a bearing on whether or not the crash is a serious one.

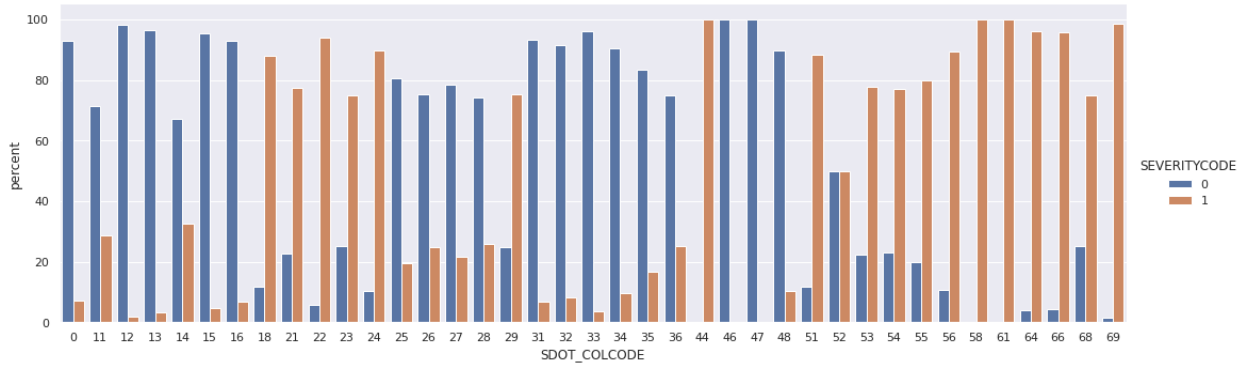
Countplot of COLLISIONTYPE by SEVERITY



Some of the categories here, e.g., ‘Left Turn’, are disproportionately serious accidents. Others, such as ‘Sideswipe’ are disproportionately not serious. Certain others, e.g., ‘Pedestrian’ are more likely to be serious than others, but engaged in additional analysis on those categories (see below).

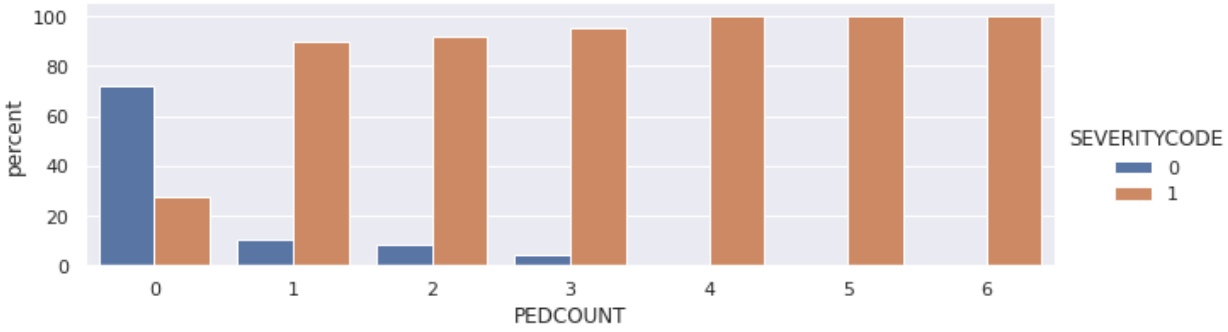
The next data point reviewed was the SDOT_COLCODE feature. The SDOT_COLCODE feature maps to a SDOT_COLDESC feature that provides a human readable description of the description. For the ease of display, the SDOT_COLCODE feature is shown in the graph below, however I used SDOT_COLDESC in the model for greater explainability.

SDOT_COLCODE (Accident description) by SEVERITYCODE



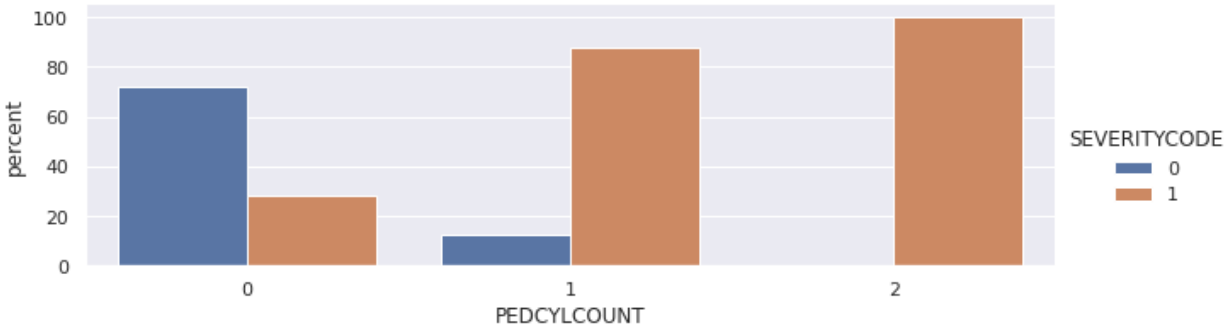
The above graph shows that many of the description codes correlate with a particular accident severity, e.g., most type ‘0’ accidents are not serious.

PEDCOUNT by SEVERITYCODE



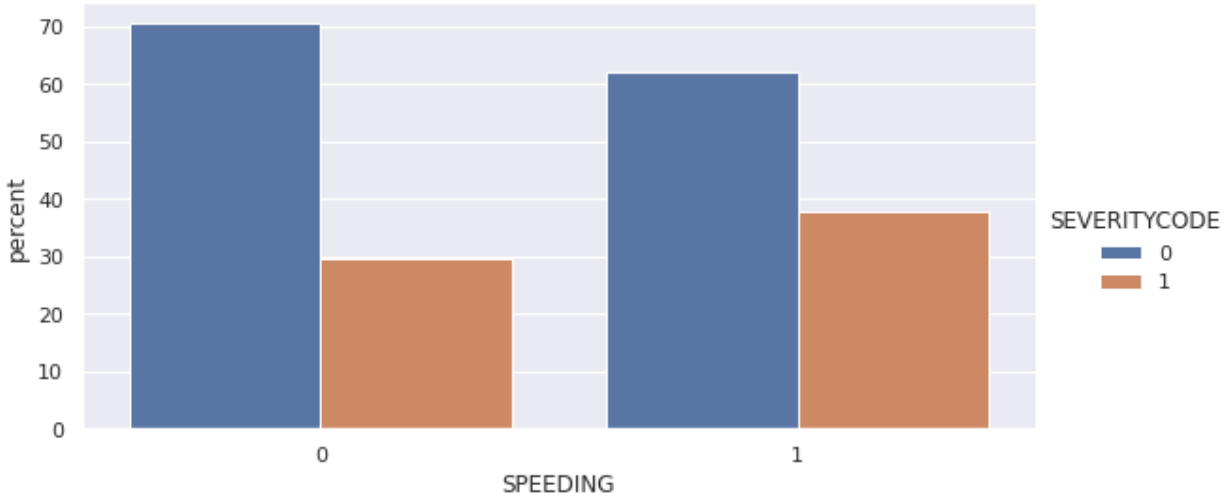
The number of pedestrians in a crash correlates pretty strongly with the likelihood of the accident being a serious one.

PEDCYLCOUNT by SEVERITYCODE



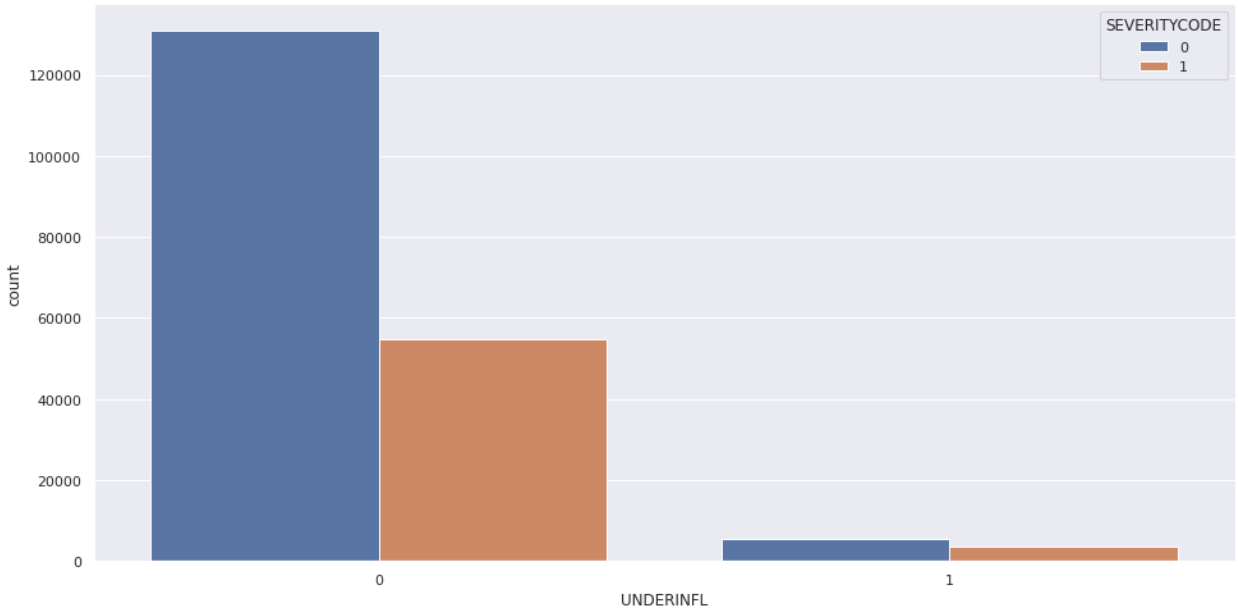
Likewise, the number of cyclists in a crash correlates strongly with seriousness. Let's look at speed.

SPEEDING indicator



Accidents where speed was a factor seem to have a slightly higher chance of being serious than when speed isn't a factor.

UNDERINFL indicator



Accidents where the driver was under the influence do seem to be disproportionately serious, but there aren't a very large number of these relative to the whole. A similar distribution was seen for INATTENTIONIND.

The most heavily correlated features with SEVERITYCODE were found to be COLLISIONTYPE, PEDCOUNT, PEDCYLCOUNT, and SDOT_COLCODE. Features that described the driver's state at the time of the accident, e.g., UNDERINFL and INATTENTIONIND provided only a minor correlation.

I attempted to engineer a 'TIME' and 'DAY OF WEEK' features from the INCDTTM feature, but further analysis showed that there was no practical correlation between either of these features and SEVERITYCODE, so I removed these.

I then transformed the categorical data values into binary features using one-hot vector encoding for the ADDRTYPE, COLLISIONTYPE, SDOT_COLDESC, WEATHER, ROADCOND, and LIGHTCOND features. I then performed further investigation on these created fields for redundancy.

I investigated the top correlations between the various features to see if the number of features could be reduced. I noticed that some of the provided data columns have an overlap with categorical data in some of the other categories, e.g., there is both a 'HITPARKEDCAR' feature, as well as a category within the 'COLLISIONTYPE' for 'Parked Car'. Likewise, I was able to reduce some of the redundant features around the pedestrian and cyclist descriptors, keeping only the features that indicated the number of pedestrians and cyclists involved, PEDCOUNT and PEDCYLCOUNT respectively, as they correlated the most with SEVERITYCODE. In a similar fashion, I was able to reduce some category values for SDOT_COLDESC that correlated with features such as PEDCYLCOUNT.

4. Discussion

4.1. Applying Standard Classification Algorithms

To start, I separated the dataset into a training and test dataset where the test data was a randomly selected 20% of the total.

In terms of model, I chose between several classification algorithms. For a structured learning problem with an imbalanced dataset, I chose a few different algorithms to test: Logistic Regression, Random Forest, and Gradient Boost. I had begun with SVM as well, but each run of the SVM algorithm on the dataset took several hours to run and did not provide a performance improvement over any of the others. For Logistic Regression, I used a class-weight of .701 to reflect the imbalanced nature of the data toward the negative class, .701 being the proportion of data belonging to the negative class.

4.2. Performances of Different Models

I have included the Jaccard Similarity score for feature for informational purposes, however the key metric selected for the analysis is the F1-score. This is because the F1-score provides a much more robust evaluation of an imbalanced dataset in terms of precision and recall. The performance of the various models appears below.

Performance of Classification Models. Best performance labeled in red.

	Logistic Regression	**Random Forest**	**Gradient Boost**
Jaccard Similarity Score	.763	.734	.734
F1 Score	.730	.652	.730
No. of True Positive	26058	27266	26186
No. of False Positives	1359	151	1249
No. of False Negatives	7855	10185	7902
No. of True Negatives	3653	1323	3606

The logistic regression model performed best overall, with the highest F1-score (tied with Gradient Boost) and the best values for false negative/true negative. The Random Forest model had the lowest F1-score, but provided the highest number of true positives and lowest number of false positives. In this problem, we care more about the false negative rate, since the consequences of encountering (or being in) a serious car accident are inherently more significant than the consequences of a false positive (e.g., changing a route to avoid the accident).

5. Conclusions

In this study, I analyzed the relationship between Seattle traffic accidents and details regarding the crash and the driver. I identified the number of pedestrians involved in the crash, the number of cyclists involved in the crash, and the type of collision as the most salient variables that predict the likelihood of a crash being a serious one. I built classification models that predict whether or not an accident is likely to be a serious one. These models can be useful in a number of ways. They can be valuable to motorists as well as logistics companies when it comes to driving strategy, as well as route selection in the event of

encountering a crash (e.g., avoid the area if a cyclist was involved in the crash). Finally, the underlying data can serve as a reminder to drivers to pay closer attention to areas where the factors for a serious crash are likely to exist (e.g., where large numbers of pedestrians or cyclists are present).