Predicting Car Accident Severity

Luigi De Marco

1 Introduction

While car accidents are typically considered to be random events, there are a large number of factors that can influence the severity of an accident. The severity of an accident can be classified according to what type of damage occurred as well as whether any injuries or fatalities were involved. The factors that can influence the severity of an accident can include, for example, weather, visibility, or whether excessive speed was involved. As such, it stands to reason that a predictive model can be used to describe the correlations between various factors and the likelihood of accidents of a given severity occurring.

The development of an accurate model to predict accident severity has significant implications for commuters and city traffic planners alike. If it is well understood which factors increase the probability of severe accidents occurring, a warning system may be put in place to alert commuters to drive particularly carefully or change travel plans. Likewise, if certain conditions are likely to cause severe accidents, city planners may shut roads and divert traffic while those conditions prevail.

Here, we build a machine learning model to predict the severity of car accidents based on a variety of factors. In what follows, we will describe the data set we will use to build such a model.

2 Data

2.1 Introduction to the Data Set

To build a model to predict the severity of accidents, we use a dataset provided by the Seattle Department of Transportation (SDOT). The raw data may be accessed online at https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv.

The data contains 194,673 records of accidents in the Seattle area from 2004 to present. Each record corresponds to an accident of a given severity. The severity of the accident is labeled according to whether property damage occurred or whether an injury was incurred. We note that this is an imbalanced data set, with property damage incidents occurring roughly three times more frequently than injury incidents.

In addition to the severity label, each record contains 37 additional attributes describing the conditions under which the accident occurred. These include numerical attributes, such as the

number of people involved in a collision and the number of vehicles involved, as well as categorical attributes, such as the weather conditions when the incident occurred and whether or not a parked car was hit. A full description of the attributes can be found online at https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf.

In order to build a model, we want to make predictions based on features that are known beforehand. That is, while the number of vehicles involved in an accident may be a good indicator of whether an accident is severe or not, this information cannot be used to develop a warning system to alert commuters. Instead, we will focus on features such as weather conditions, road conditions, and light conditions to make predictions *before* severe accidents occur.

	SEVERITYCODE	INCDTTM	ADDRTYPE	WEATHER	ROADCOND	LIGHTCOND
0	2	3/27/2013 2:54:00 PM	Intersection	Overcast	Wet	Daylight
1	1	12/20/2006 6:55:00 PM	Block	Raining	Wet	Dark - Street Lights On
2	1	11/18/2004 10:20:00 AM	Block	Overcast	Dry	Daylight
3	1	3/29/2013 9:26:00 AM	Block	Clear	Dry	Daylight
4	2	1/28/2004 8:04:00 AM	Intersection	Raining	Wet	Daylight

The above table shows an example of the raw data that will be used to build a model. Each incident is labeled by a severity code ('SEVERITYCODE') that indicates whether only property damage occurred (label 1) or whether injury was involved (label 2). In addition, each record contains attributes describing the conditions of the incident. In the above example, these attributes include: the date and time of the incident ('INCDTTM'), the location type at which the incident occurred ('ADDRTYPE'), the weather conditions ('WEATHER'), the road conditions ('ROADCOND'), and the light conditions ('LIGHTCOND'). Intuitively, we expect these to be good predictors of accident severity. In the next section, we will analyze the features more carefully and determine which may give rise to the best model.