Description of the Data

This study aims to develop a model that correctly identifies the severity of the traffic accidents in the city of Seattle. To this end, a relatively rich road accident records dataset will be analyzed by using several different machine learning models.

Data Source

The dataset is taken from Kaggle and it is originally hosted by the City of Seattle. It includes accident records with all types of collisions and covers accidents between the years of 2004 and 2020. It is updated weekly.

Data Description

The copy of the data that is used for this study has 38 columns, all of which include different features, and 194673 entries.

Methodology

The data set is presented in one csv file and has relatively large number of features, namely 38 columns. Before working on a model, the dataset was studied in detail by understanding every feature. The main principals of exploratory data analysis and data preprocessing were followed accordingly and will be briefly explained below.

Developing a machine learning model that predicts the severity of road traffic accidents with a relatively high rate of accuracy would be certainly beneficial for everyone. Moreover, since machine learning models, with complex algorithms enable researchers to identify salient factors that contribute to accidents and their severity. Such factors are sometimes not easily discernable among large datasets.

Exploratory Data Analysis

In a preliminary step before working on a model, I analyzed first the general structure of the data with types of variables, their meanings, the relation of several feature with our target feature accident severity, which is coded as “SEVERITYCODE” in the dataset.
Severity has two categories in our model:

1: Property damage

2: Injury

Using various visualizing techniques, I intended to see if there was a significant, consistent seasonal pattern.

Feature Engineering and Data Preparation

After the initial exploratory data analysis, for feature engineering and data preparation, 7 features, that seemed most relevant for our inquiry have been selected. Feature selection is an important step in data analysis to develop an efficient model that generates high percentage of accuracy. Including too many features in a model might adversely affect the model performance and also makes it harder to design a deployable model that could be used effectively in practice. Selected features are shown in table below:

	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND	COLLISIONTYPE	UNDERINFLUENCE	SPEEDING
0	2	Overcast	Wet	Daylight	Angles	N	N
1	1	Raining	Wet	Dark - Street Lights On	Sideswipe	N	N
2	1	Overcast	Dry	Daylight	Parked Car	N	N
3	1	Clear	Dry	Daylight	Other	N	N
4	2	Raining	Wet	Daylight	Angles	N	N

For a better understanding a feature was renamed and missing values were with different values in accordance with the characteristic of that specific feature.

For categorical variables label encoding one-hot encoding and “get_dummies” methods were applied.