Predicting Car Accident Severity

Applied Data Science Capstone Project
IBM Data Science Professional Certificate

Mustafa Bal

1. Introduction

1.1 Background

As one of the leading international institutions that has set road traffic accidents as one of its primary agendas on global scale, the World Health Organization (WHO) reports that roughly 1,35 million people die every year as result of road traffic accidents (WHO, 2020). Road traffic accidents also leave between 20 - 50 million people injured, some of whom suffer long term health problems and some physical disability or impairment. According to the Center for Disease Control and Prevention, In the United States, where approximately 10 million individuals are involved in a car accident annually (Rolison et al., 2018), car accidents are the number one cause of death for teenagers. Unfortunately, the numbers are not different for other countries. According to the records of the WHO, injuries that inflicted as a result of traffic accidents are the leading cause of death for children and young adults between the ages of 5 and 29.

Road traffic accidents result in also significant economic losses for states and for the individuals and their families who are involved in accidents. The WHO estimates that the cost of road traffic accidents to most countries to be approximating 3% of their gross domestic product. We cannot also ignore the long-term social consequences of traffic accidents on especially children and families. In brief, as a serious global problem road traffic accident should be on the agenda of researchers.

1.2 Problem Statement

As a serious global problem road traffic accident has been on the agenda of researchers as well as governments. I believe to tackle with this issue effectively we should develop a coordinated and multi-disciplinary approach. In this respect, this study would like to contribute to the analysis of the road traffic accident problem by employing tools and methodologies of Data Science field.

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 was analyzed by using several different machine learning models.

1.3 Interest

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.

2. Data

2.1 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.

2.2 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.

3. 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.

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

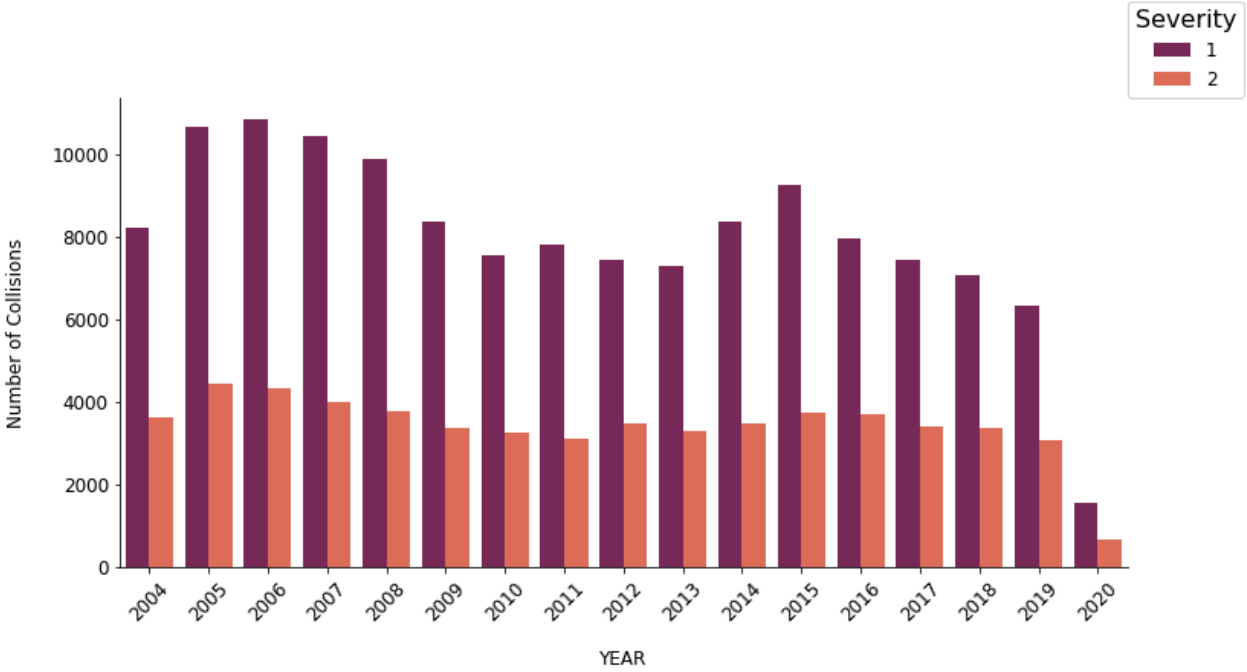


Figure 1: Yearly Distribution of Accidents

As the table above shows, the number of accidents and severity level of accidents decreased from 2005 to 2013. After a slight increase in 2015, it has decreasing trend again.

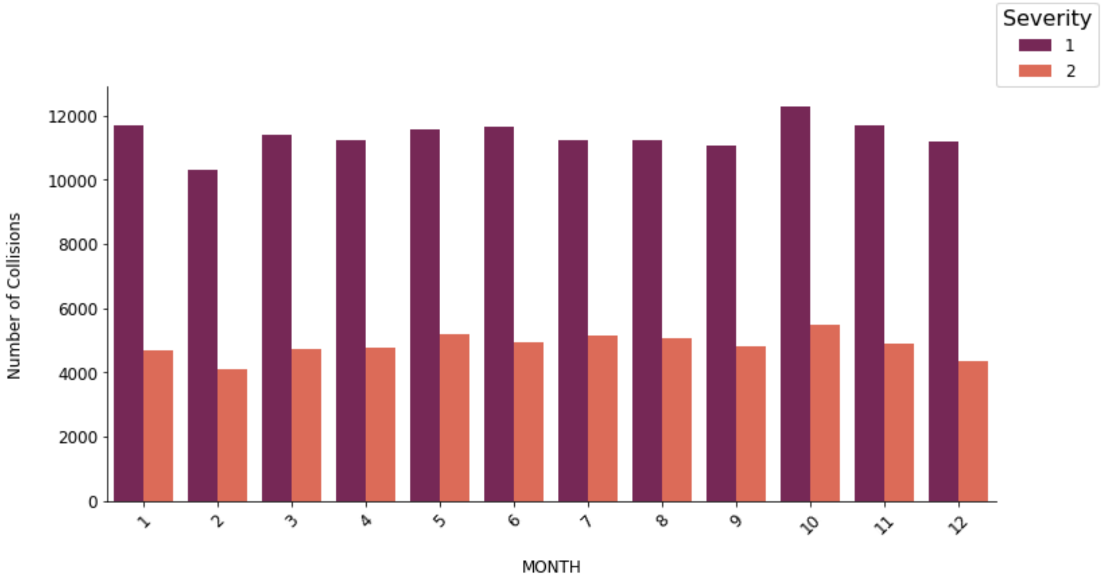


Figure 2: Monthly Distribution of Accidents

The month of October has slightly higher number of accidents.

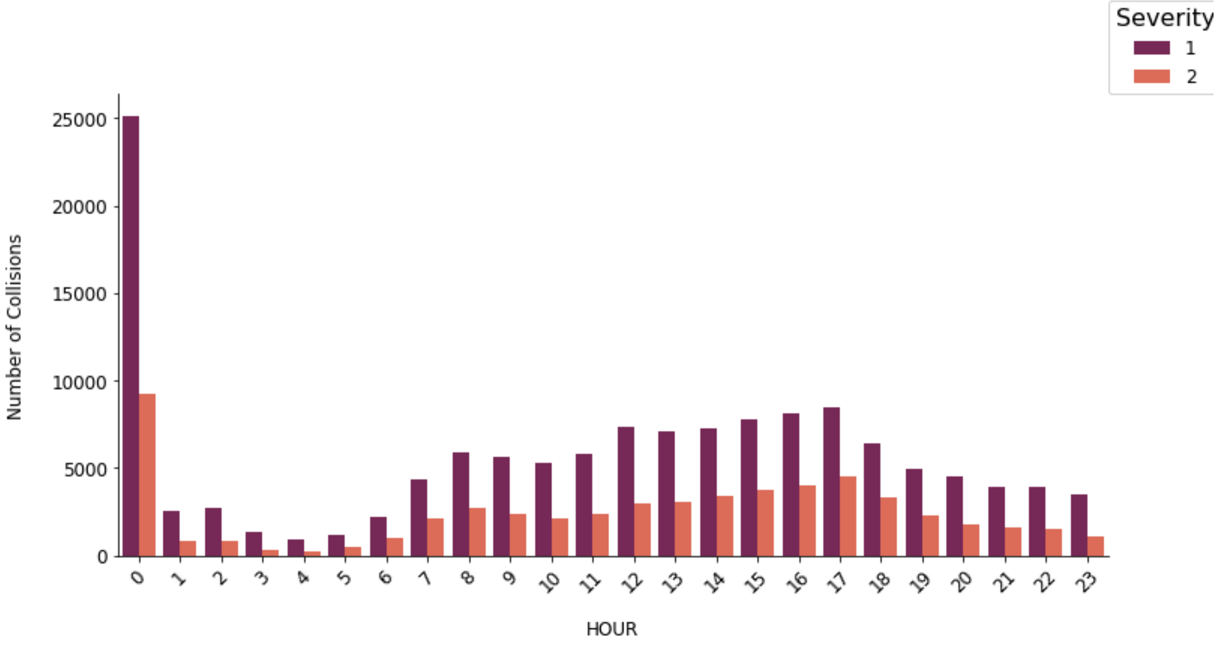


Figure 3: Distribution of Accidents on Hourly Basis

In terms of hourly basis, the initial analysis shows that the majority of accidents take place between 08:00 and 17:00 hours. It is also clear from the table that highest number of accidents occur at midnight.

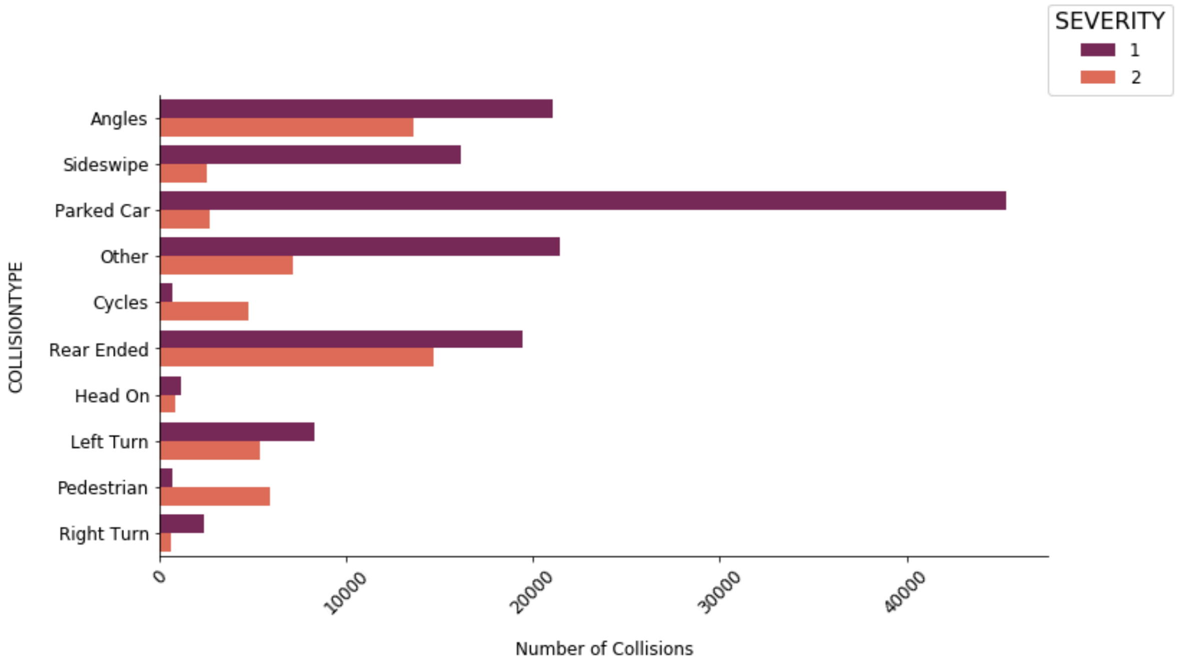


Figure 4: Severity Levels by Collision Types

“Parked Car” category is the most frequent collision type but in terms of severity, “Angles” and “Rear Ended” types of accidents have the highest number of severity category of 2.

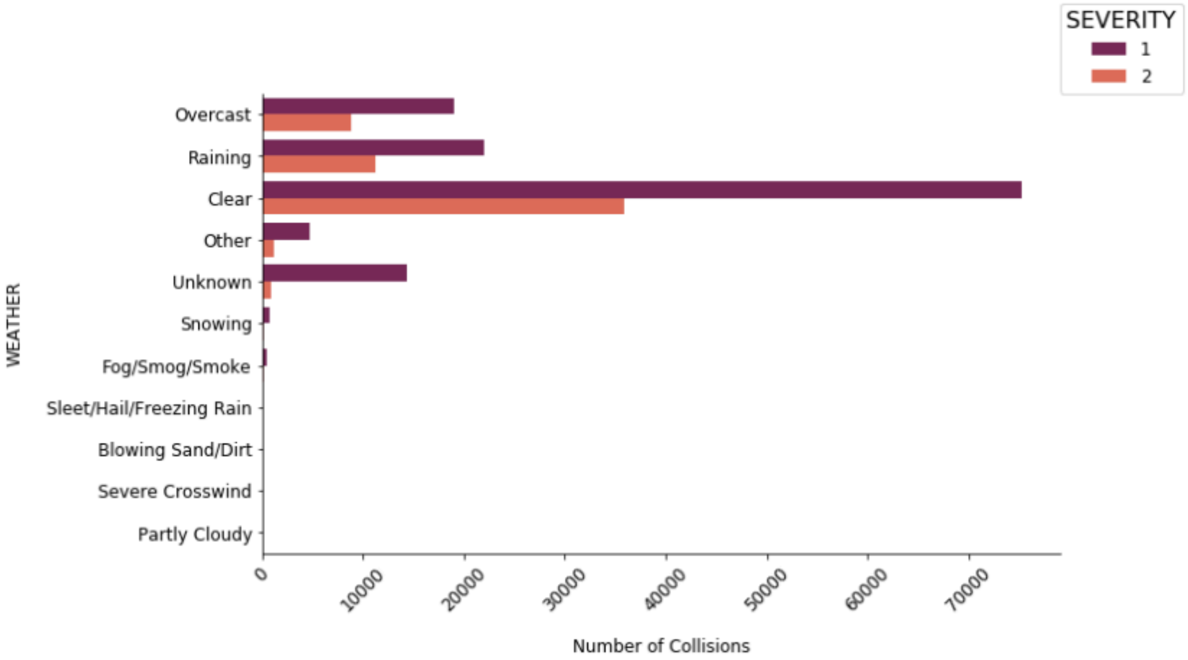


Figure 5: Severity Levels by Weather

Most accidents occur in clear weather conditions.

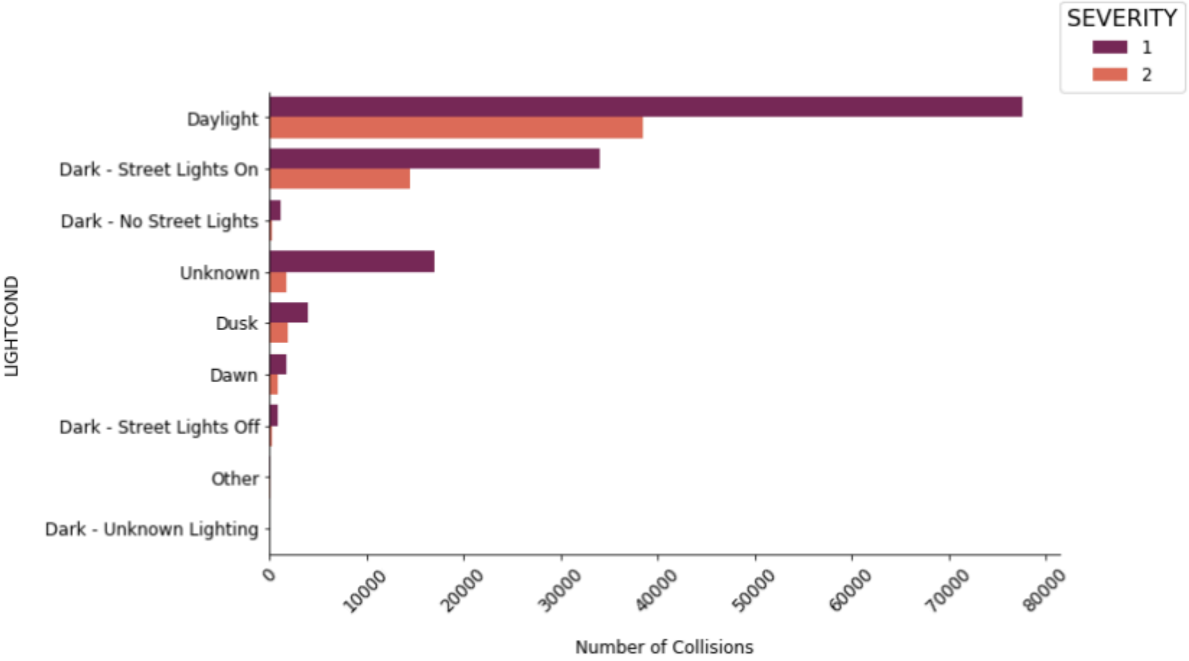


Figure 6: Number of Collisions by Light Conditions

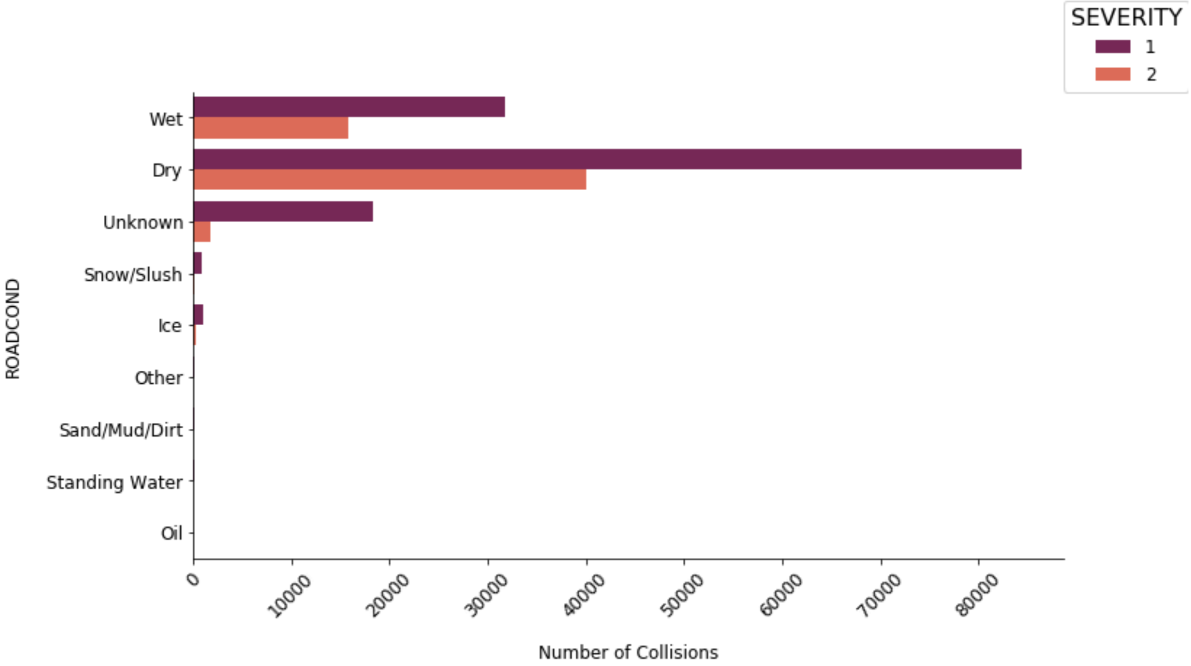


Figure 7: Number of Collisions by Road Conditions

Most accidents happen in daylight, under clear weather, and dry road conditions but these three variables alone have little explanatory value.

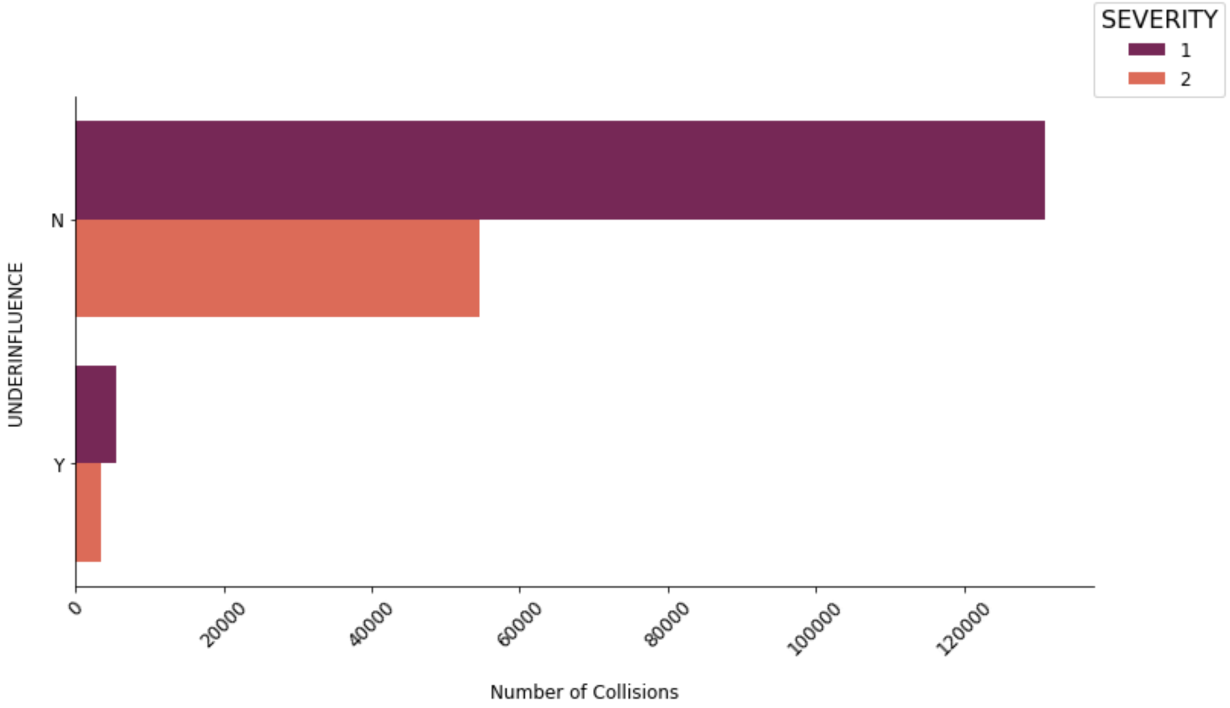
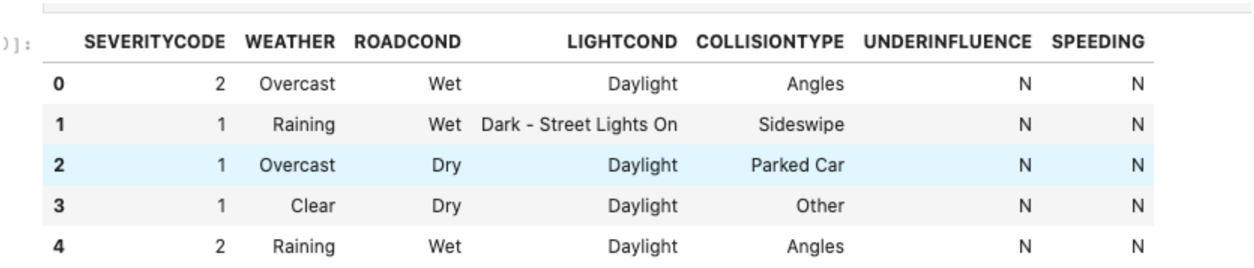


Figure 8: Number of Collisions Under Influence

The number of accidents under influence are relatively small in comparison to general total number of accidents.

3.2 Feature Engineering and Data Preparation

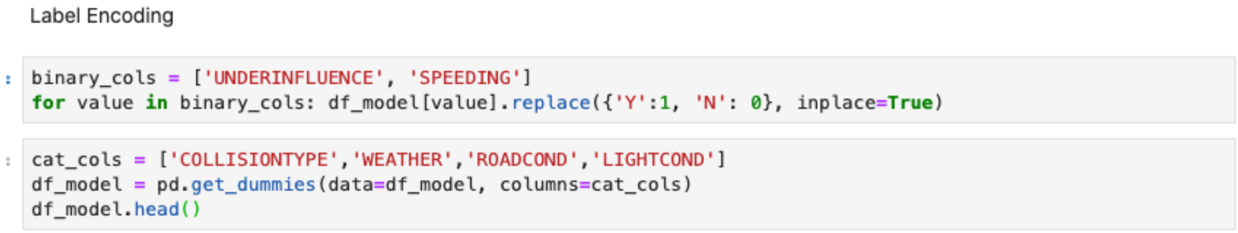
After the initial exploratory data analysis, for feature engineering and data preparation, 7 features, that seemed most relevant for our inquiry have been selected. Feature selection is an important step in data analysis to develop an efficient model that generates high percentage of accuracy. Including 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.



Label Encoding

```
binary_cols = ['UNDERINFLUENCE', 'SPEEDING']
for value in binary_cols: df_model[value].replace({'Y':1, 'N': 0}, inplace=True)
```

```
cat_cols = ['COLLISIONTYPE','WEATHER','ROADCOND','LIGHTCOND']
df_model = pd.get_dummies(data=df_model, columns=cat_cols)
df_model.head()
```

4. Model Development

As the famous saying “There is No Free Lunch in Data Science” suggests that there is no one model that works best for every problem. Although it some models are believed to be performing better for certain types of problems, it is hard to determine the best option without first implementing different models. Given these considerations, 6 different machine learning algorithms were used for problem.

1. Logistic Regression
2. K-Nearest Neighbour (KNN)
3. Decision Tree
4. Random Forests
5. Gradient Boosting (GBM)
6. Support Vector Machine (SVM)

5. Results

The final comparison of models shows that SVM has the highest accuracy score by a very small margin. However, given its high F1 and accuracy score and its relatively easier implementation, I personally prefer using Random Forest model.

Model	F1 Score	Accuracy
Decision Tree	0.685978	0.749634
KNN	0.697810	0.740979
LOGISTIC REGRESSION	0.696848	0.749994
GRADIENT BOOSTING	0.691827	0.750995
RANDOM FOREST	0.692660	0.750970
SVM	0.691684	0.751098

Conclusion

In this study, different machine learning models were used to analyze and better understand road traffic accidents. As the final analysis shows, various models can predict the severity of an accident with an approximate accuracy of 75%. This score is quite significant and would certainly be useful in authorities' approach to better understand and identify the contributing factors of road traffic accidents and their severity. Predicting the severity of an accident correctly could also be helpful for first responders to react and work more efficiently.

Furthermore, such analyses with Data Science methods and machine learning models also measure the quality of the gathered data. In return, in coordination with the relevant authorities, the evaluation of data quality might provide further insights for creating better information or data templates that intend to gather more relevant data in road accident records.

References:

Rolison, Jonathan & Dorchin-Regev, Shirley & Moutari, Salissou & Feeney, Aidan. (2018). What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions, and road accident records. Accident Analysis & Prevention. 115. 10.1016/j.aap.2018.02.025.

WHO, (October 5, 2020). Road traffic injuries. World Health Organization.

https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

Data Source: https://www.kaggle.com/jonleon/seattle-sdot-collisions-data

Metadata: https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf