CAR ACCIDENT SEVERITY

BUSINESS UNDERSTANDING (PROBLEM STATEMENT)

According to the Global status report on road safety, conducted by World Health organization in 2013, based on the information gathered on road safety from 182 countries, accounting for almost 99% of the world's population, the total number of road traffic deaths are unacceptably high at 1.24 million per year.

Every day, many accidents occur with various degree of severity based on different external and internal factors. Therefore, it is important to define the factors leading the severity of accidents and to generate a model which predicts the severity type of accidents best. Based on these developed models, it is possible to help the authorities take some actions to reduce the risks.

The purpose of this study is to try to understand the factors that contribute to the severity factor of a vehicle collision in the city of Seattle by using the data from the city of Seattle recorded between the years 2004 and 2020. In order to predict the severity of the collision based on various factors such as road and wet conditions in dataset, different machine learning algorithms were developed. Based on the understanding the impact of different factors on the collision outcome, it is hoped to provide meaningful insights into how to prevent such collisions so that the drivers can be alerted in advance. After Exploratory Data Analysis and Data Cleaning, predictive models were developed and evaluated by different metrics

DATA

The data for this project was provided by IBM's Applied Data Science Capstone Project. This dataset includes many features to build a robust predictive machine learning algorithm. This data contains all types of collisions which occurred between the years of 2004 and 2020 in the city of Seattle. It has a total of 194673 records and 38 different attributes.

METHODOLOGY

After careful exploratory data analysis, some variables were selected that researcher think, drive the severity level of accidents. For this undertaking, the following variables were selected to build a robust machine learning model.

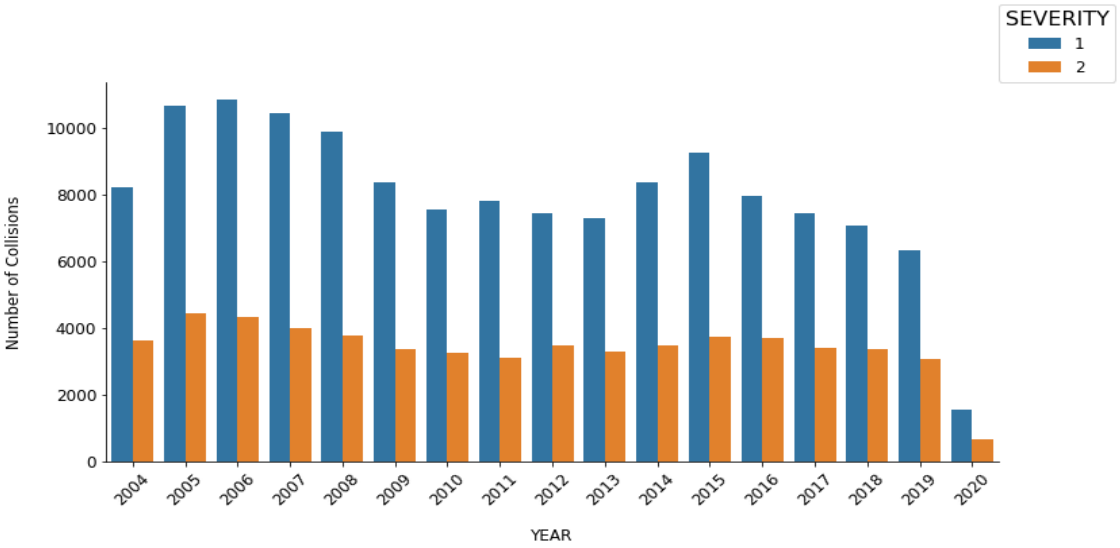
	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND	COLLISIONTYPE	INATTENTIONIND	UNDERINFL	SPEEDING
0	2	Overcast	Wet	Daylight	Angles	NaN	N	NaN
1	1	Raining	Wet	Dark - Street Lights On	Sideswipe	NaN	0	NaN
2	1	Overcast	Dry	Daylight	Parked Car	NaN	0	NaN
3	1	Clear	Dry	Daylight	Other	NaN	N	NaN
4	2	Raining	Wet	Daylight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

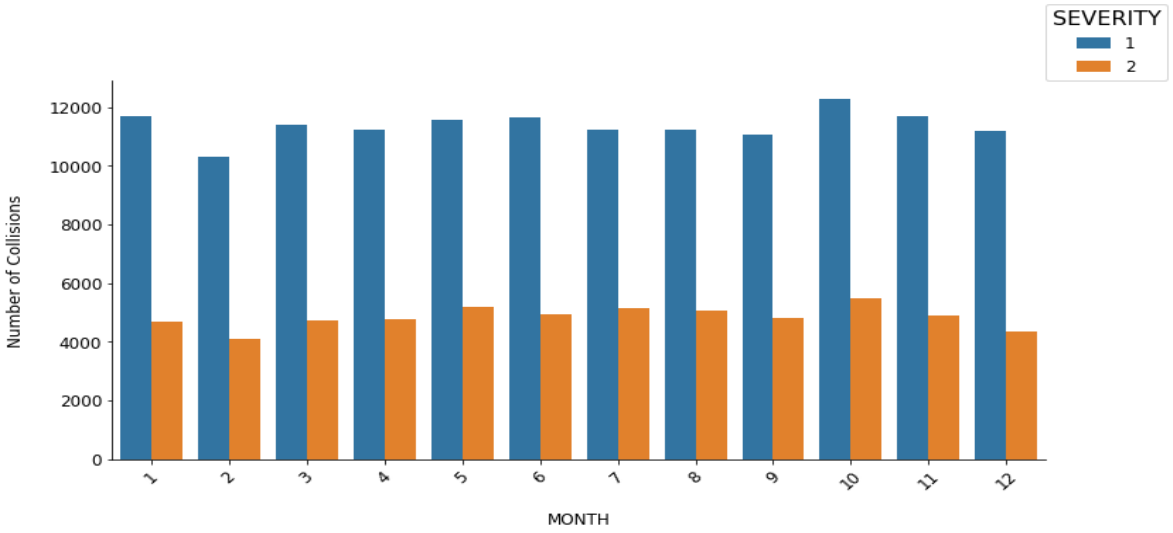
The target variable of the study is the SEVERITY CODE, which is used to measure the severity level of an accident.

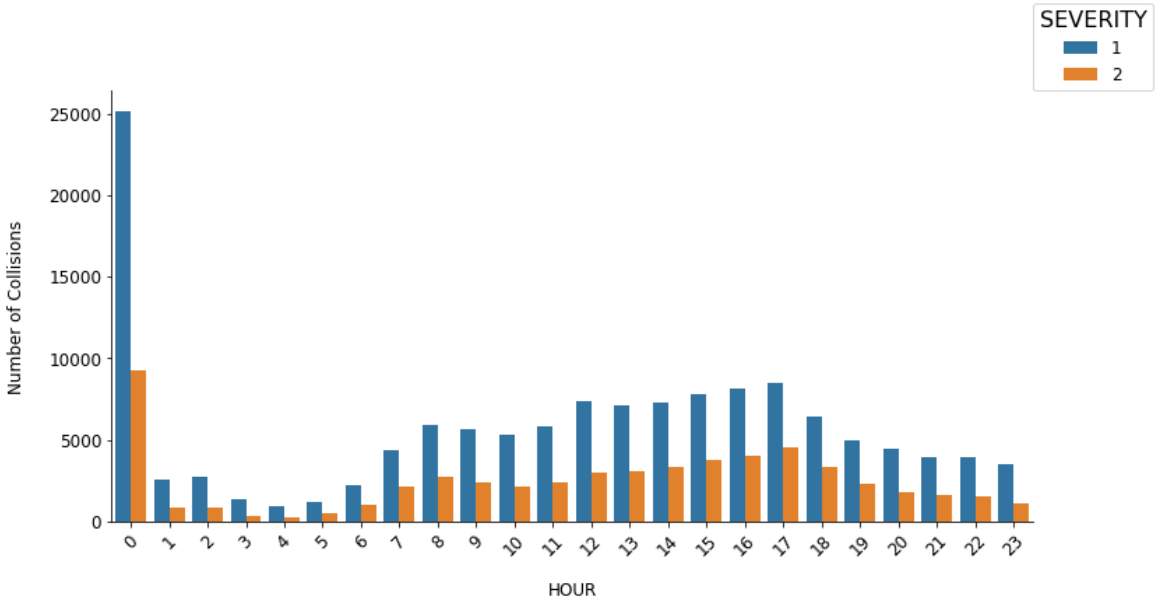
EXPLORATORY DATA ANLAYSIS

1. Severity Levels by Year & Month & Hour

   Following tables illustrates the severity level of accidents by the time of year, month, and hour.

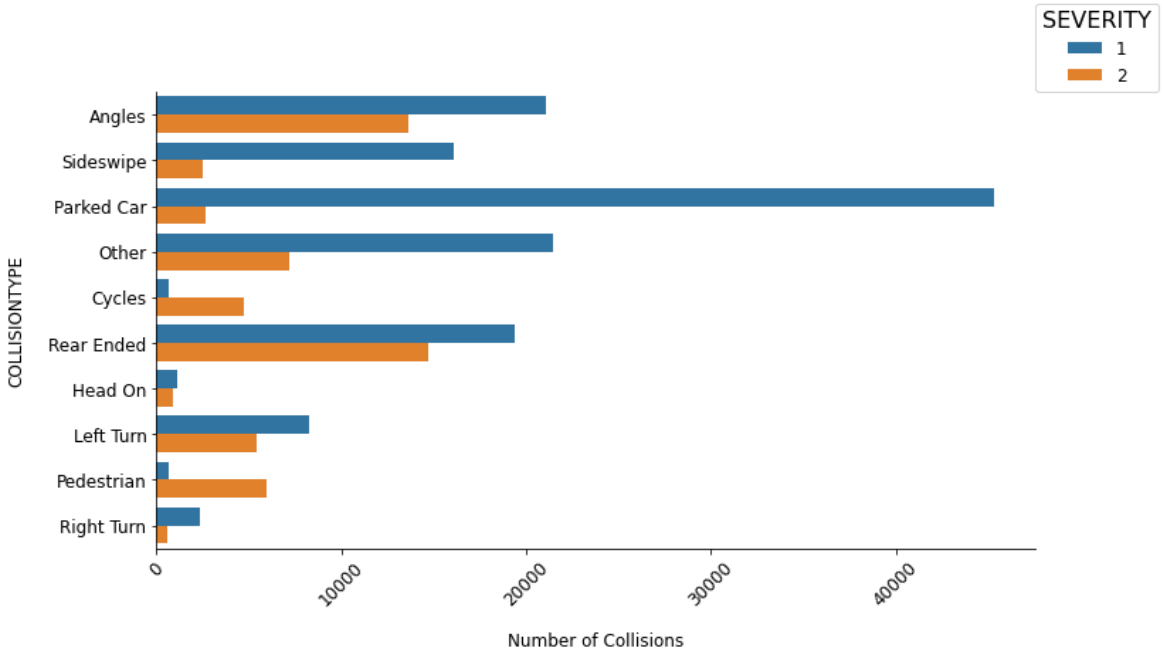

SEVERITY
1
2
Number of Collisions
0
2000
4000
6000
8000
10000
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
YEAR


SEVERITY
1
2
Number of Collisions
0
2000
4000
6000
8000
10000
12000
1
2
3
4
5
6
7
8
9
10
11
12
MONTH


SEVERITY
1
2
Number of Collisions
0
5000
10000
15000
20000
25000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
HOUR

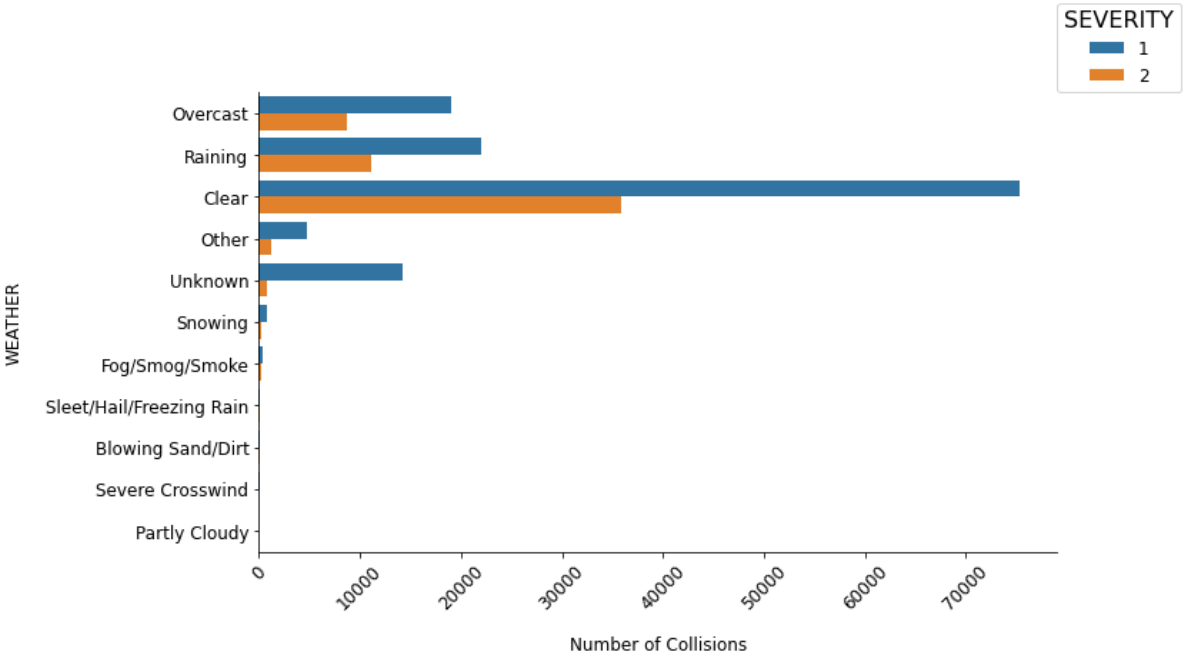
The number of accidents for both severity classes have been decreasing from the year of 2005 to 2013. Then it seems increasing to the year of 2015, and then decrease again. The month of October has slightly higher than rest of the months in terms of severity level of accidents for both types. The highest number of accidents happen at midnight.

2. Severity Levels by Collision Types



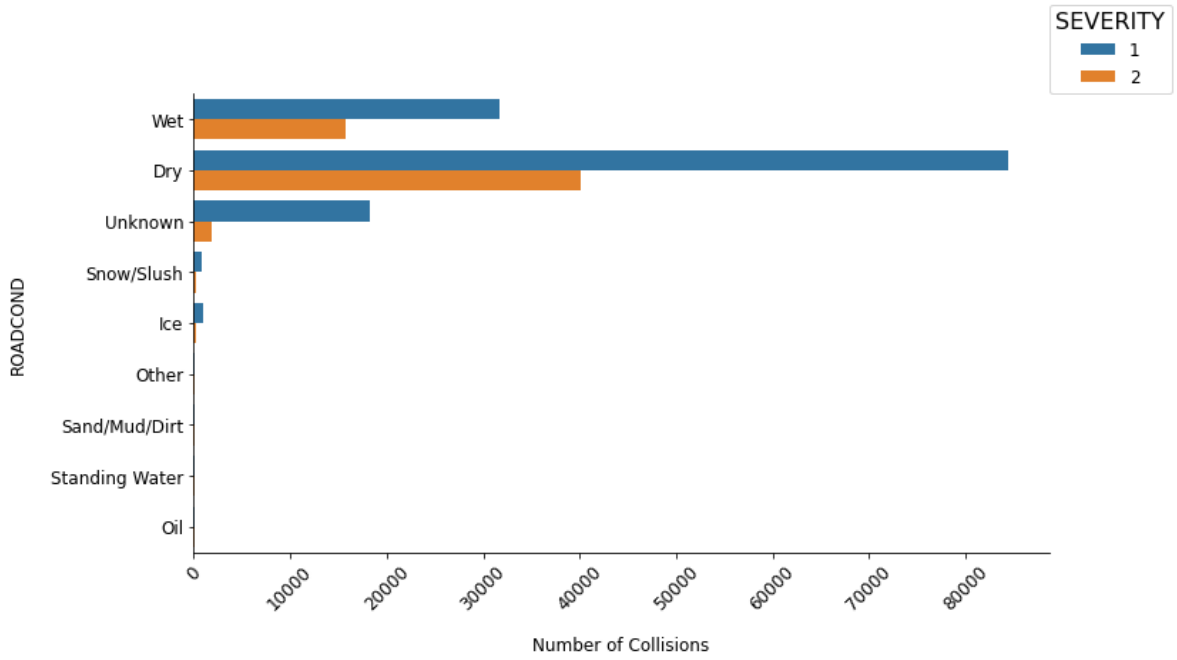
It is interesting to note that

3. Severity Levels by Weather Conditions



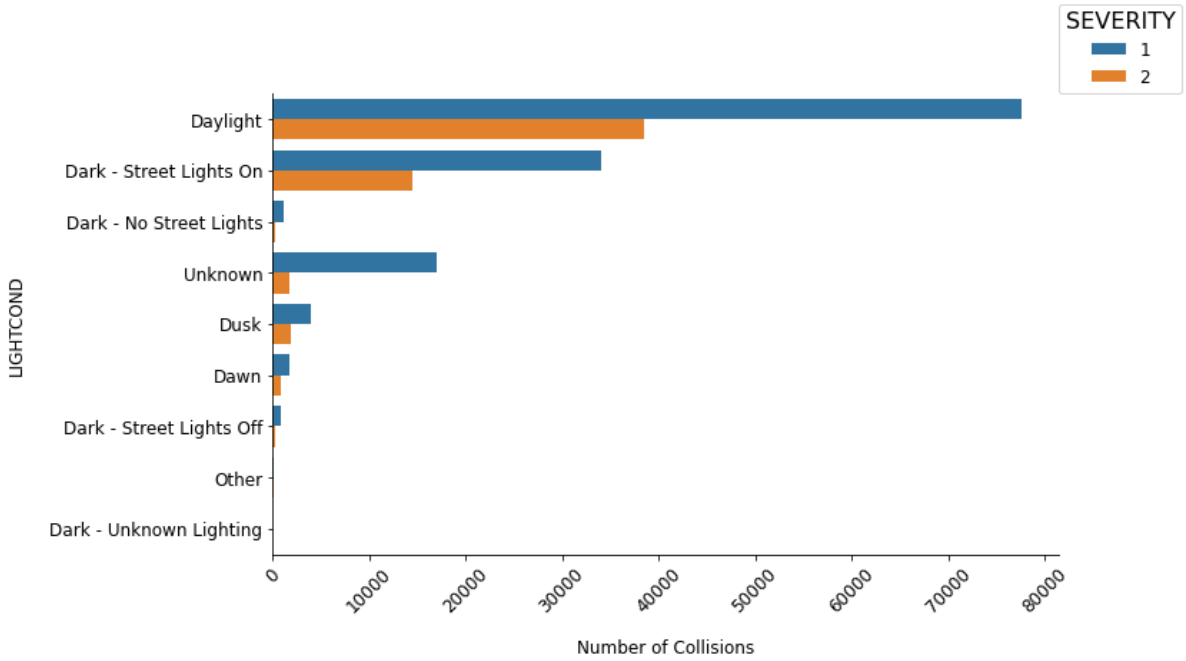
Majority of accidents occur at clear weather conditions even though it is more for severity type 1, which is injury collision. Rainy and overcast are other important weather conditions for severity levels of accidents.

4. Severity Levels by Road Conditions



It seems that accidents occur in the presence of dry and wet conditions. Snowy and Icy road conditions accounts for small fraction of accidents.

5. Severity Levels by Light Conditions



It is interesting that most accidents occur in daylight, and then dark but streetlights are on.

MODEL DEVELOPMENT

Five different classification models (supervised machine learning algorithms) were developed to classify the data.

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

MODEL EVALUATION AND COMPARISON

	Model	F1 Score	Accuracy
0	Decision Tree	0.683414	0.748658
1	KNN	0.687520	0.729678
2	Logistic Regression	0.692575	0.747939
3	Gradient Boosting	0.688820	0.749120
4	Random Forests	0.683676	0.748761

When evaluating these different models by F1 and accuracy score, even though the scores are very similar one another, Gradient Boosting Model performs the best to predict the severity levels of accidents based on the selected independent variables.

DISCUSSION & CONCLUSION

The selected features from the dataset have been used to classify the severity of the accidents. Five different machine learning algorithms were developed, namely, Decision Tree, KNN, Random Forests, Logistic Regression, and Gradient Boosting. Based on the accuracy scores, the Gradient Boosting algorithm offers the best performing model with an accuracy score of 0.7491. Exploratory data analysis indicates that the most vehicle accidents occur in a clear and dry road during the day. Therefore, it is important to emphasize that Seattle Transportation Agency should emphasis more on the drivers' training in terms of accidents.