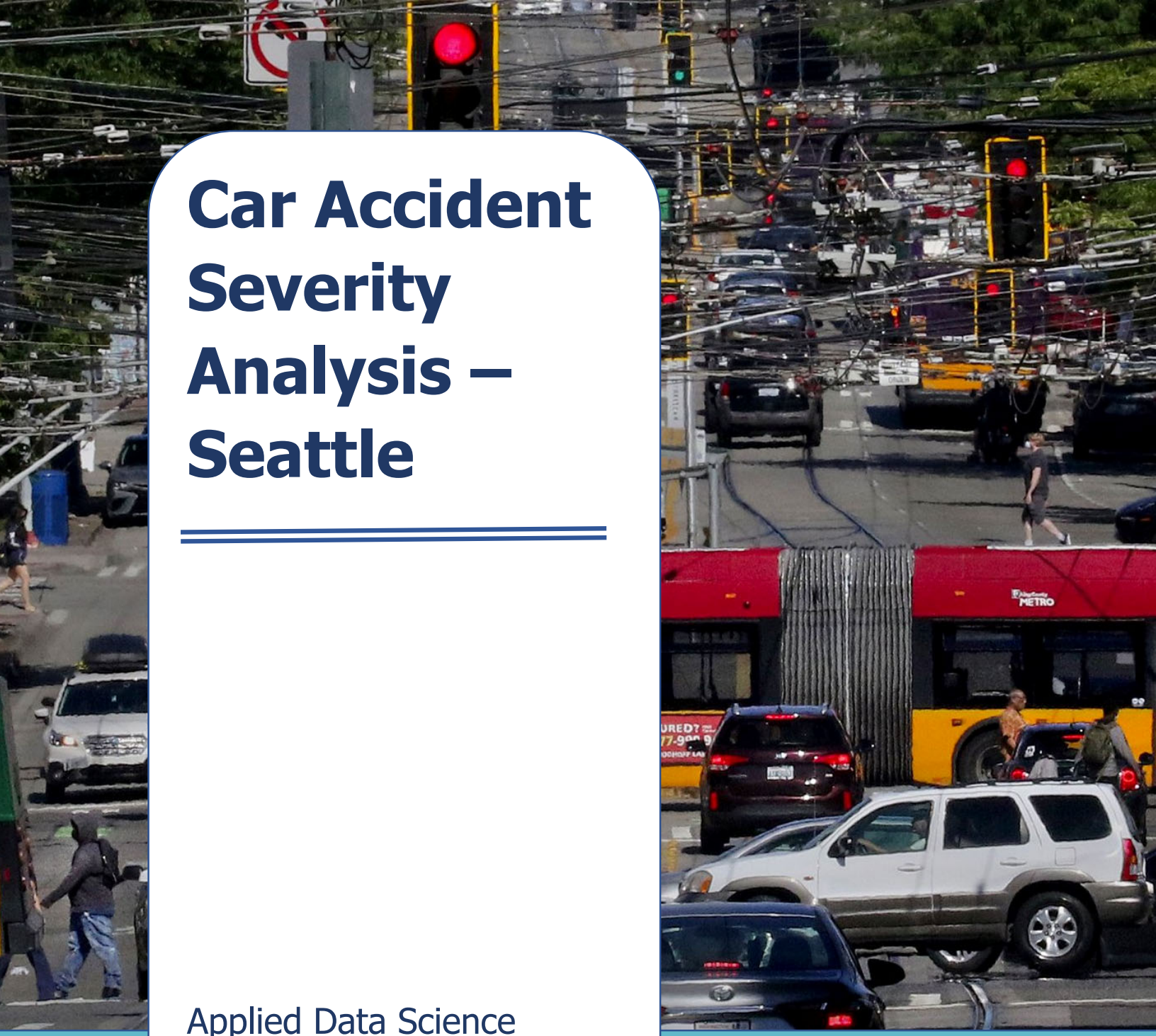


Car Accident Severity Analysis – Seattle

Applied Data Science Capstone Project

Submitted By:

Kirk Xyrrus Villanueva

October 2, 2020

Table of Contents

Introduction

Background

From January 2004 to May of 2020, there have been over 190 thousand recorded cases of car accidents in Seattle alone. These include various types of vehicles such as Cars, Trucks, Motorcycles, Bicycles, and others.

The estimated total number of road traffic is 1.35 million every year or one death every 24 seconds. Almost half of these deaths are coming from those with the least protection such as motorcyclists, cyclists, and pedestrians. On a global scale, road traffic injuries and crashes are estimated to be the eighth leading cause of death across all ages.

This analysis aims to determine the factors or attributes that might be contributing to these accidents and predict the severity level.

Stakeholders

This analysis will help the Seattle residents to be aware of the attributes which may lead to accidents related to road accidents and how they can address the issues to lessen the risk of these accidents.

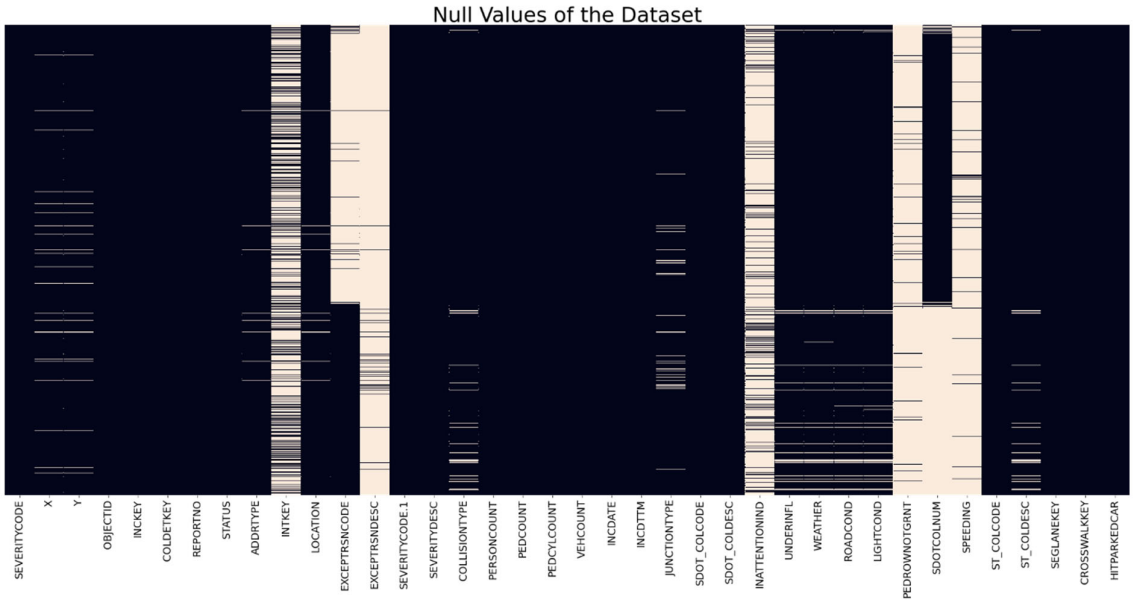
Understanding the Data

Data Source

The data used for this analysis is from the Traffic Records Group of the Seattle Department of Transportation, Traffic Management Division. This includes all types of collisions at the intersections and mid-blocks of a segment from January 2004 to May 2020.

Data Cleansing

The dataset used for this analysis is a CSV file comprised of 194,671 rows and 38 columns. The data requires a lot of preprocessing and data cleansing due to a lot of null and inconsistencies in some of the attributes.



The table shows the null occurrences in the dataset. Some of the columns will not be utilized as the values are unique and will skew the model.

All the rows with null and '***Other***' values are removed from the dataset as they don't add value to the analysis.

Feature Selection

The dependent variable for this analysis is the **SEVERITYCODE** with two possible values:

1 – Property Damage Only Collision

2 – Injury Collision

The independent variables for this analysis to classify and determine the severity of the accidents are the following:

ADDRTYPE – Flags whether the collision happened in an **alley**, a **block** or, an **intersection**.

WEATHER – Weather condition during the time of the collision.

ROADCOND – Condition of the road during the collision.

LIGHTCOND – Light condition during the collision.

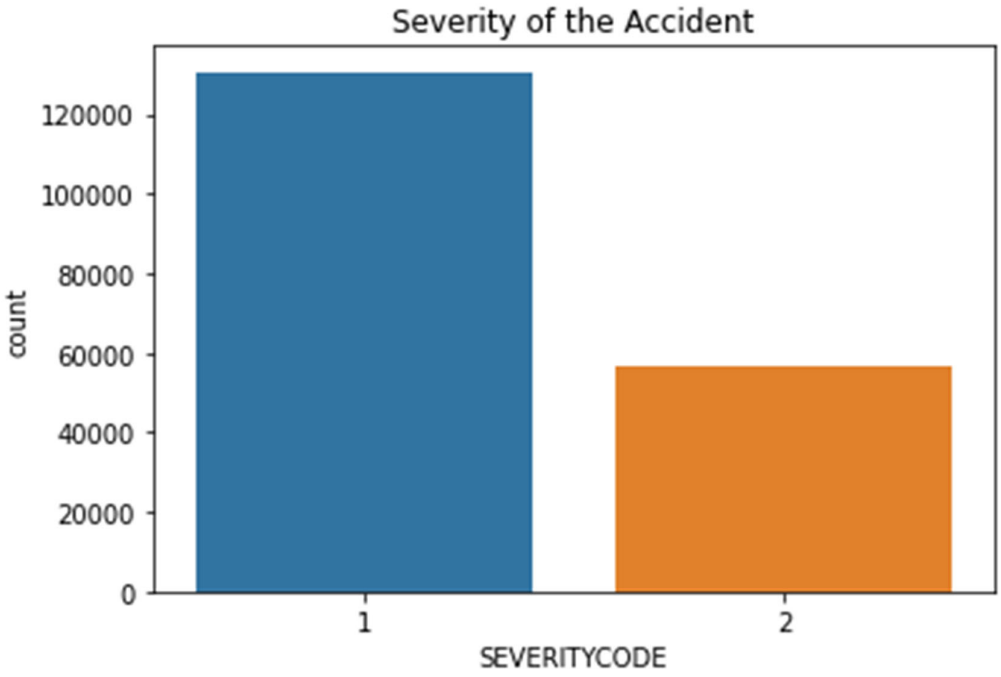
UNDERINFL – Whether the driver of the vehicle was under the influence of drugs or alcohol.

Methodology

Exploratory Data Analysis

DEPENDENT VARIABLE

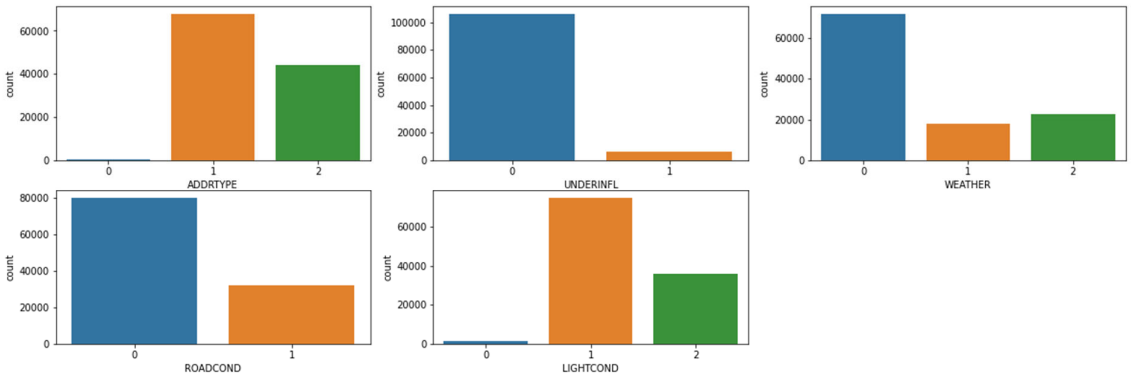
SEVERITYCODE



The distribution of the target variable (SEVERITYCODE) shows that there is an imbalance between the possible values. After removing all the rows with null or 'Other' values, there are 169,247 rows left in the dataset. Only 33% or 55,542 rows belong to injury collision making the dataset imbalanced. To address this, a new dataset will be created including all rows with a SEVERITYCODE value of **2** and the same number of rows with SEVERITYCODE value of **1** will be randomly pulled from the original dataset. The new dataset will have a total of 113,766 rows.

INDEPENDENT VARIABLES

As part of data processing, the categorical text data were converted to numerical data through the ***LabelEncoder*** function for the model to understand the data.



ADDRTYPE

This variable has three possible values: Alley (0), Block (1), and Intersection (2).

UNDERINFL

This attribute has two possible values: N (0) and Y (1). For consistency, N and Y values are replaced with 0 and 1 respectively.

WEATHER

This variable has three possible values: Clear (0), Overcast (1), and Raining (2). Some of the minor observation categories were grouped with the major ones:

- *Partly Cloudy* was binned with **Clear**
- *Sleet/Hail/Freezing Rain*, *Raining*, and *Snowing* are grouped under **Raining**
- *Overcast*, *Blowing Sand/Dirt*, *Fog/Smog/Smoke*, and *Severe Crosswind* are grouped under **Overcast**

ROADCOND

This attribute has two possible values: Dry (0) and Slippery (1). Grouping was also used in this feature:

- *Wet*, *Ice*, *Oil*, *Snow/Slush*, and *Standing Water* are all grouped under **Slippery**
- *Dry* and *Sand/Mud/Dirt* are grouped under **Dry**

LIGHTCOND

This variable has three possible values: Dark (0), Daylight (1), and Limited Visibility (2). Grouping was also used in this variable:

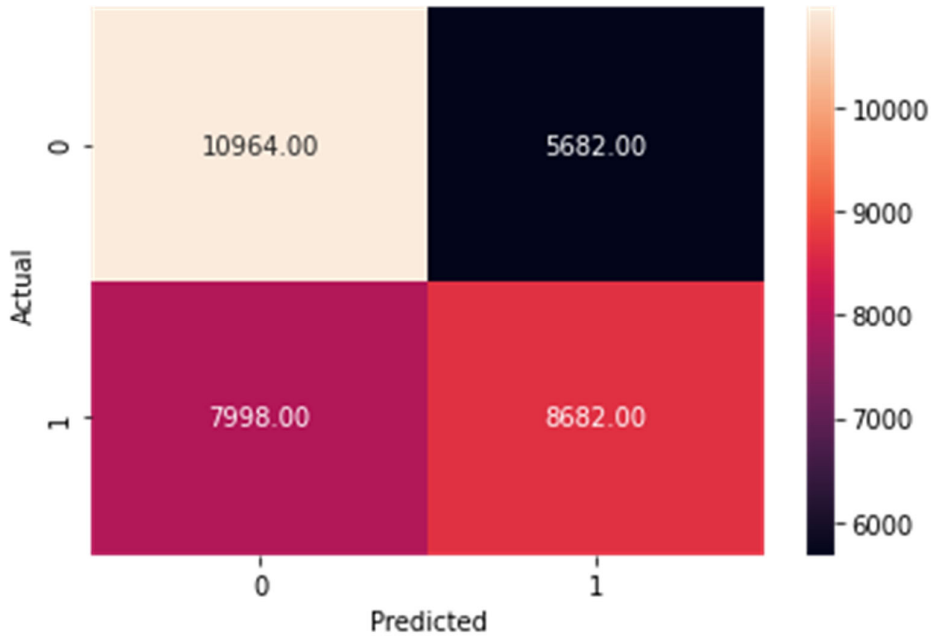
- *Dark – No Street Lights*, *Dark – Street Lights Off*, and *Dark – Unknown Lighting* are grouped under **Dark**
- *Dark – Street Lights On*, *Dawn*, and *Dusk* are grouped under **Limited Visibility**

Model

To predict the severity of the accident in this analysis, Decision Tree and Logistic Regression models will be used. SEVRITYCODE will be used as the dependent variable (y) and all other columns in the new dataset will be the independent variables (X). The data will be split to train and test datasets. 70% of the data will be used as the Training Set where the models will be built. The remaining 30% will serve as the Test Set to check the accuracy of the models.

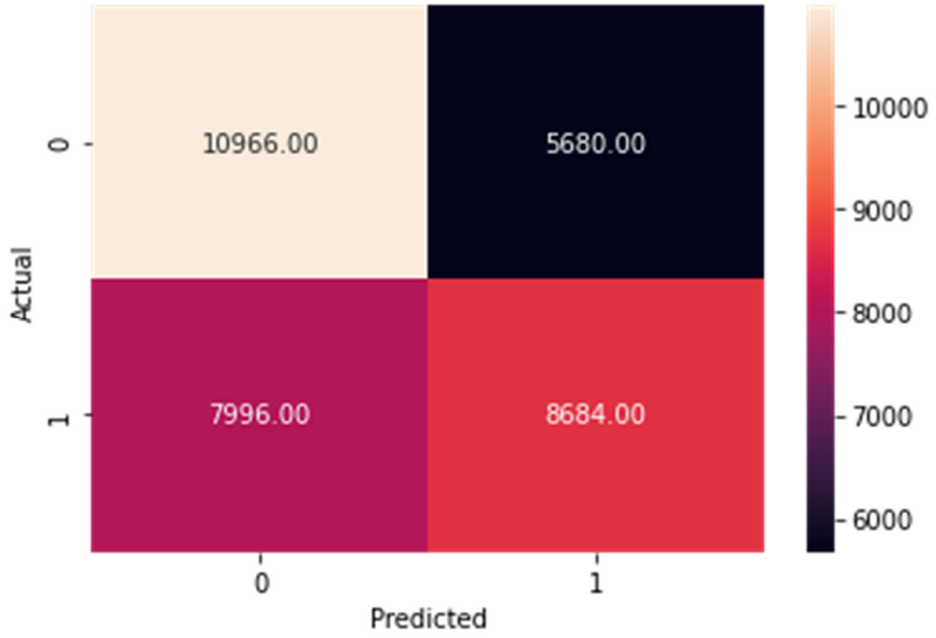
Results

Decision Tree



The criterion used for this model is **gini** with a max depth of **5**. The f1-score, or the accuracy (highest value is ***1.0***), for this model is **0.59**. Recall and Precision are 0.66 and 0.58 respectively.

Logistic Regression



The results for the Logistic Regression model is very close to that of the Decision Tree Model. The solver used in the model was **liblinear** using parameter C or the inverse of regularization equal to **0.1**. The f1-score for this model is **0.59** and the logloss was. Recall and Precision are 0.66 and 0.58 respectively.

Conclusion

Both models have performed almost identical concerning f1-score, recall score, and precision score. Therefore, either model can be used to predict the severity of the car crashes in Seattle. But given the poor quality of data, the accuracy of the models is not performing well.

Recommendations

With the results of the analysis, the primary recommendation is for the SDOT to have a better standard of data gathering and encoding. Should the base data have less null and unusable data, the analysis could have arrived with a better prediction and higher accuracy.

Another recommendation, given that there are cases wherein the accidents transpired in a low visibility category, the local government should focus on the areas without street lights and make developmental projects to install light posts to lessen the risk of car crash-related incidents.

References

https://www.who.int/gho/road_safety/mortality/en/

https://www.seattle.gov/transportation/about-us

https://projects.seattletimes.com/2020/traffic-lab-data-page/