Table of Contents

1. Introduction

1.1 Background

The whole world is looking for a way to prevent car accidents from happening in their area car accidents can cause a several of problems to both governments and individuals Fatalities and car jamming are part of the problems added on that the cost of health care systems

As stated by https://www.cdc.gov/injury/features/global-road-safety/index.html 1.35 million people are killed due to car accidents about 3700 a day which make car accidents the eighth leading cause of death

1.2 Problem

Seattle is like any other place in the world suffer from car accident and according to collision data 30% Of mischance leads to harm collision which is close to 1/3 of all car accidents this considered a tall number. our mission is to try to know the pattern of these accident and if we can predict what cause the high count of severities.

In most of the cases known accidents are caused by drivers not paying considerations on the street, manhandling drugs and liquor or by driving in tall speeds. shockingly, these human blunders cannot be anticipated nor be calculated since of that we'll attempt to utilize other components like climate, street condition, and visibility

1.3 Stakeholders

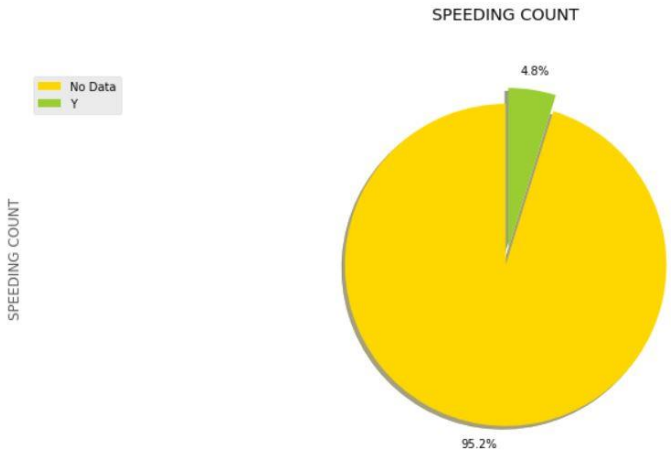
The intended interest group of this project is Seattle government, police, salvage gatherings, and to wrap things up, vehicle protection foundations. The model and its outcomes will give some guidance to the intended interest group to settle on smart choices for diminishing the quantity of accidents and severity

2. Understanding Data

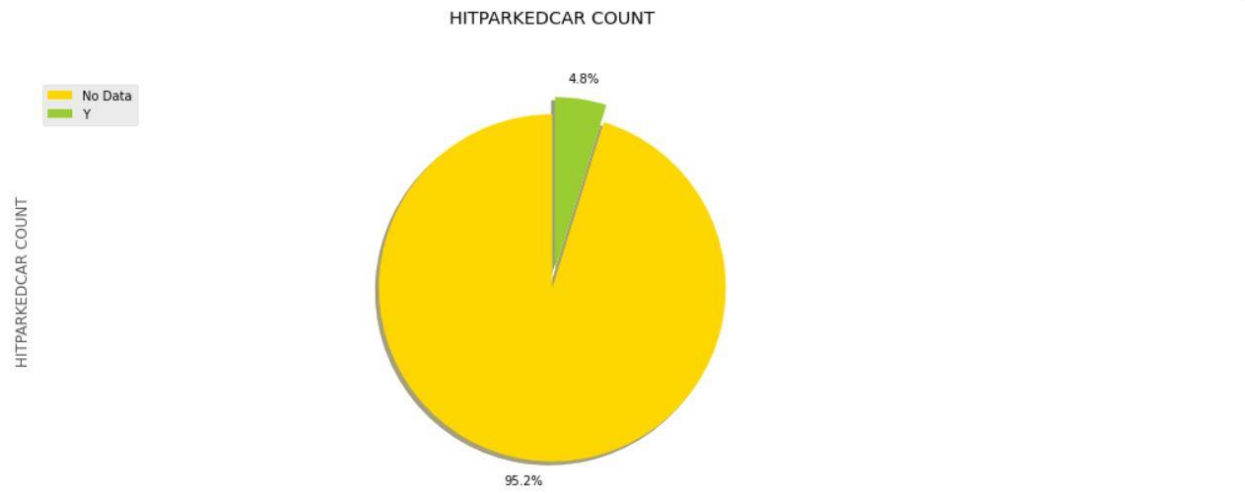
2.1 Data Cleaning

There is a great deal of issues in the informational collection. The dataset has complete perceptions of 194673 record with variety of records that can be used. Above all else, the all-out dataset was high variety in the lengths of pretty much every segment of the dataset. The dataset had a great deal of void segments which could have been useful if the information had been available there. These segments included speed or not, cross walk key, hit parked car,

In order to proceed we have first to clean our data as mentioned we have a lot null values in valuable columns like weather, road condition, and light condition (visibility) we will drop the rows that contains null values in this columns we will also drop the rows that contains any unknown data like (unknown and others) we will also replace the text into integers
In weather we will replace Clear by 0 Blowing Sand Dirt Fog/Smog/Smoke by 2 Overcast by 3 Partly Cloudy by 4 Raining by 5 Severe Crosswind by 6 Sleet/Hail/Freezing by 7 Snowing by 8
road condition we will replace Wet by 0 Dry by 1 Snow/Slush by 2 Ice by 3 Sand/Mud/Dirt by 4 Standing Water by 5 Oil by 6
light condition we will replace Daylight by 0 Dark - Street Lights On by 1 Dark - No Street Lights by 2 Dusk by 3 Dawn by 4 Dark - Street Lights Off by 5 Dark - Unknown Lighting by 6
address type we will replace Intersection by 0 Block by 1 Alley 2
we tried to use SPEEDING but we find that a lot of fields are null only 4.8 % have data wich mean we cannot use it as predictor



we tried to use HIT PARKED CAR but we find that a lot of fields are null only 4.8 % have data which mean we cannot use it as predictor



2.2 Feature Selection

A total of 4 features were selected for this project along with the target variable being Severity Code.

Feature Variables	Description
WEATHER	Weather condition during time of collision
ROADCOND	Road condition during the collision
LIGHTCOND	Light conditions during the collision
ADDRTYPE	Place of collision
JUNCTIONTYPE	where conflicting traffic flows meet

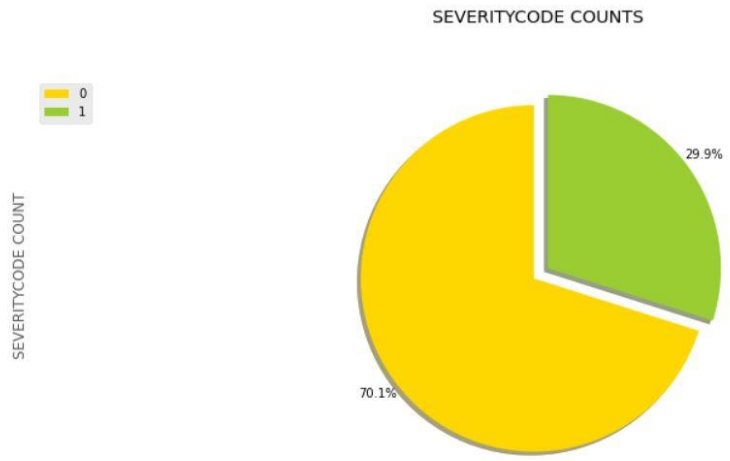
3. Methodology

3.1 Data Collection

The dataset utilized for this venture depends on auto collisions which occurred inside the city of Seattle, Washington from the year 2004 to 2020. This information is in regards to auto crashes the seriousness of every auto collision alongside the time and conditions under which every mishap happened

3.2 Exploratory Analysis

After exploring we found that are some are irrelevant and some of the data are missed so the first thing that we do is to clean the data from null values and we will also clean the data form values like Unknown and other since even if we use them as predictor we will not have any meaningful result of them
After deleting all the un useable data we found that the data is biased and the are a big difference between severity codes 29.9 to 70.1 because of that we decided to balance



These are the number before balancing

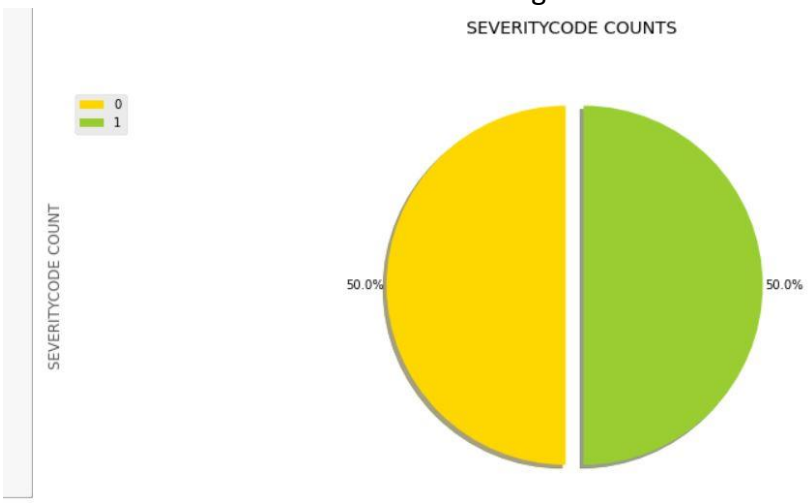
```
df1['SEVERITYCODE'].value_counts()
```

```
1    112015
2     55320
Name: SEVERITYCODE, dtype: int64
```

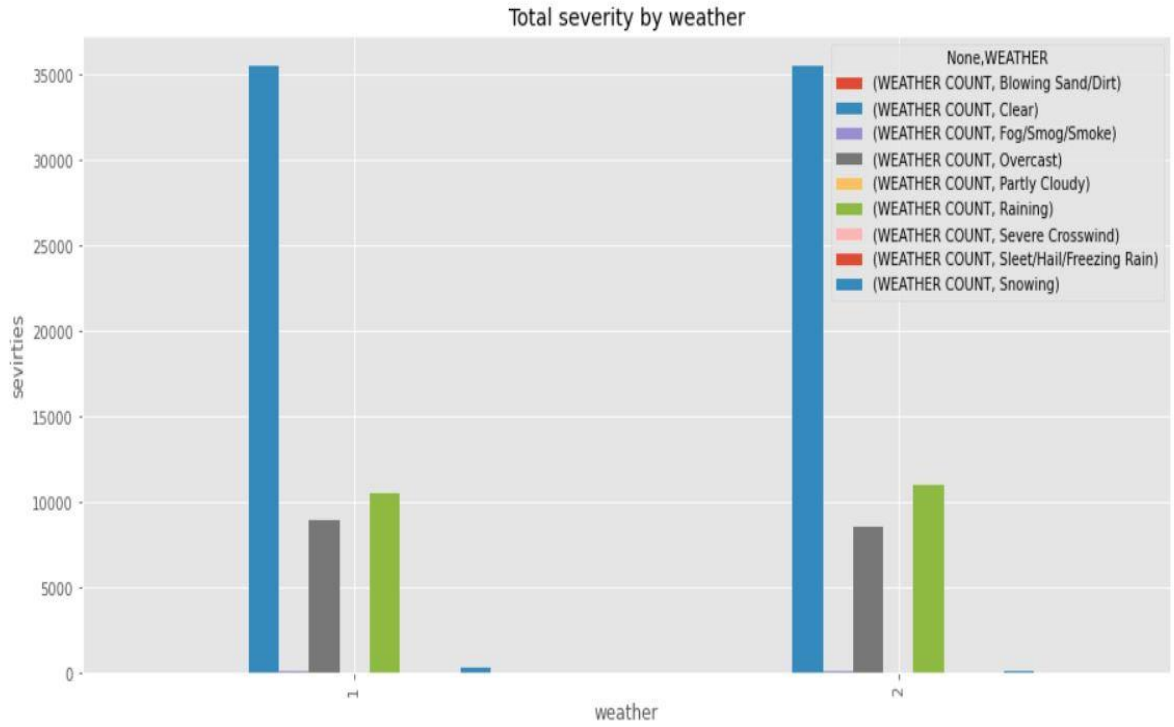
Numbers after balancing

SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND	ADDRTYPE	JUNCTIONTYPE	SEVERITYCODE COUNT
1	55320	55320	55320	55320	55320	55320
2	55320	55320	55320	55320	55320	55320

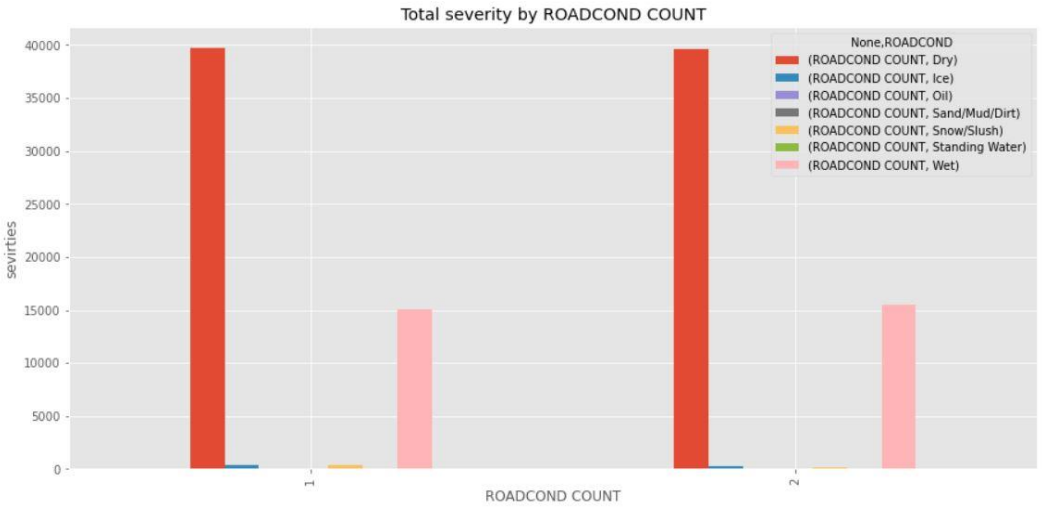
And these are the number after balancing the data we can find that they are 50/50 distributed



After balancing we wanted to visualize the data in a way that we know how each feature effects the result of the severities



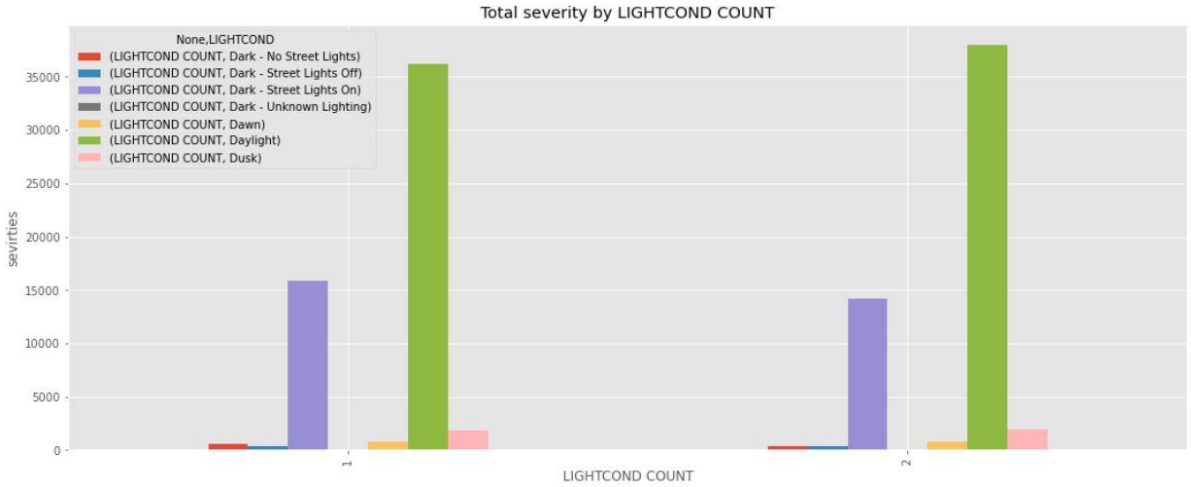
In the case of weather, we found that the results are very close to each other that mean the we cannot rely full on it



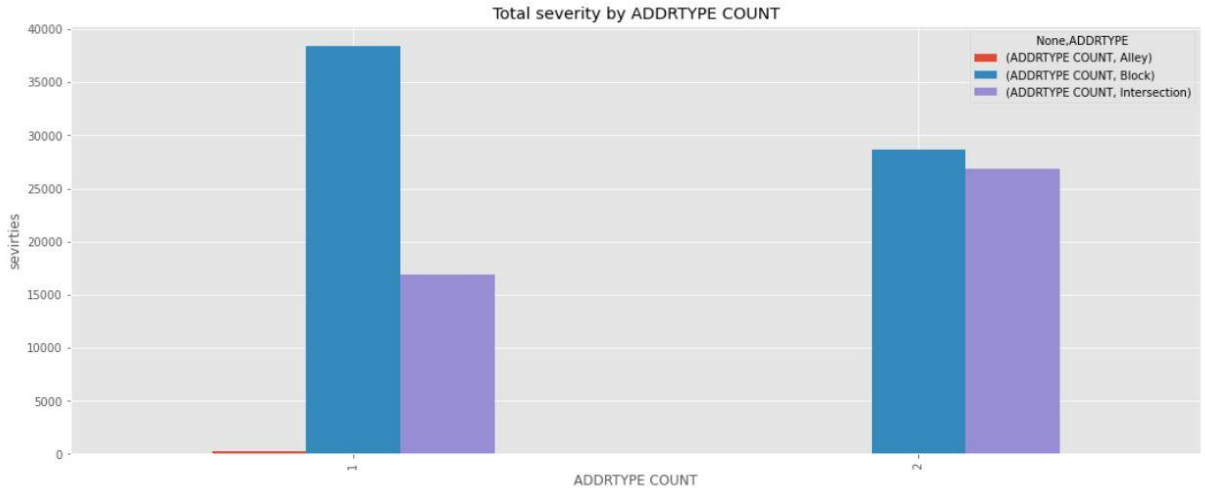

```
df1['SEVERITYCODE'].value_counts()
```

```
1    112015
2     55320
Name: SEVERITYCODE, dtype: int64
```

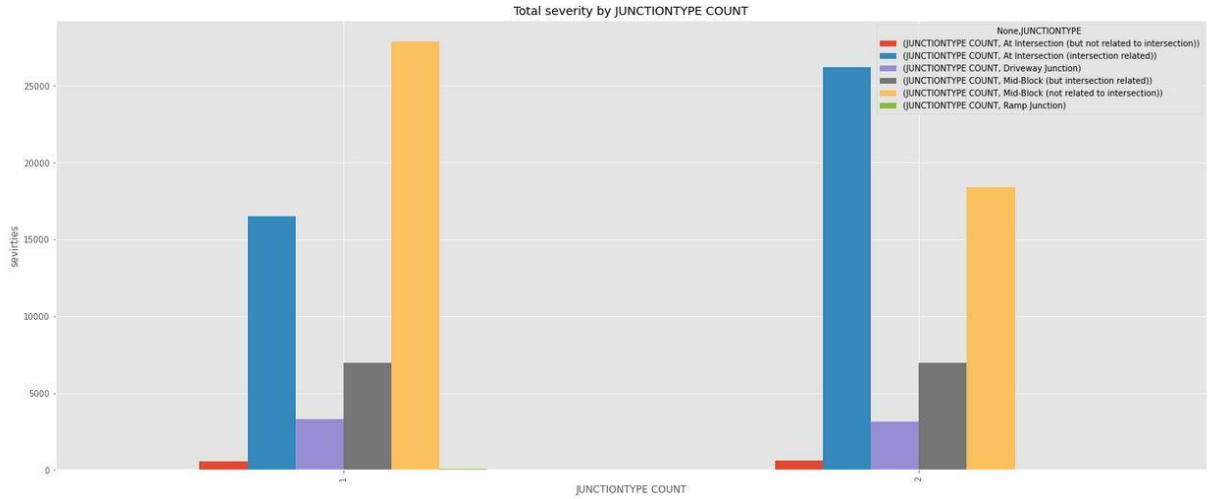
In the case of ROAD CONDITION, we found that the results are very close to each other that mean the we cannot rely full on it



In the case of LIGHT CONDITION, we found that the results are very close to each other that mean the we cannot rely full on it



In the case of address type, we found that the results are slightly different from each other that mean the we cannot rely full on it



In the case of Junction type, we found that the results are slightly different from each other that mean the we cannot rely full on it

3.3 Machine Learning Model Selection

We will try different model to train and we will choice the best on oh them

- k-nearest neighbors
- NVM
- Logistic Regression
- Decision-Tree Classifier

4. Results

4.1 k-nearest neighbors

We are going to use the famous iris data set for our KNN example. The dataset consists of five attributes: WEATHER, ROADCOND, LIGHTCOND, ADDRTYPE, JUNCTIONTYPE. The task is to predict the class to which these severities belong. There are two classes **Injury Collision**, **`Property Damage Only Collision`**. We will set k=9. train set accuracy: 0.5730409436008677 test set accuracy: 0.574295010845987

	precision	recall	f1-score	support
1	0.50	0.37	0.42	11019
2	0.50	0.63	0.56	11109
accuracy			0.50	22128
macro avg	0.50	0.50	0.49	22128
weighted avg	0.50	0.50	0.49	22128

4.2 SVM

We are going to use the famous iris data set for our SVM example. The dataset consists of five attributes: WEATHER, ROADCOND, LIGHTCOND, ADDRTYPE, JUNCTIONTYPE. The task is to predict the class to which these severities belong. There are two classes **Injury Collision**, **`Property Damage Only Collision`**. We will use kernel='rbf'. train set accuracy: 0.5911175885755604 test set accuracy: 0.5876265365148229

	precision	recall	f1-score	support
1	0.50	0.52	0.51	11019
2	0.50	0.48	0.49	11109
accuracy			0.50	22128
macro avg	0.50	0.50	0.50	22128
weighted avg	0.50	0.50	0.50	22128

4.3 Logistic Regression

We are going to use the famous iris data set for our Logistic Regression Classifier example. The dataset consists of five attributes: WEATHER, ROADCOND, LIGHTCOND, ADDRTYPE, JUNCTIONTYPE. The task is to predict the class to which these severities belong. There are two classes **Injury Collision**, **Property Damage Only Collision**. We will use C=2. train set accuracy: 0.5886546456977585 test set accuracy: 0.5841467823571945

	precision	recall	f1-score	support
1	0.57	0.69	0.62	11019
2	0.61	0.48	0.54	11109
accuracy			0.58	22128
macro avg	0.59	0.58	0.58	22128
weighted avg	0.59	0.58	0.58	22128

4.4 Decision Tree

We are going to use the famous iris data set for our Decision Tree Classifier example. The dataset consists of five attributes: WEATHER, ROADCOND, LIGHTCOND, ADDRTYPE, JUNCTIONTYPE. The task is to predict the class to which these severities belong. There are two classes **Injury Collision**, **Property Damage Only Collision**. We will use criterion="entropy", max_depth = 6. train set accuracy: 0.5014280549530007 test set accuracy: 0.5871475054229935

	precision	recall	f1-score	support
1	0.59	0.58	0.58	27640
2	0.59	0.59	0.59	27680
accuracy			0.59	55320
macro avg	0.59	0.59	0.59	55320
weighted avg	0.59	0.59	0.59	55320

5. Discussion

Algorithm	Jaccard	F1-score
KNN	0.342225	0.566901
Decision Tree	0.413015	0.587134
SVM	0.424762	0.587402
LogisticRegression	0.451674	0.579692

According to the above table we make the decision to use SVM model since it have the highest F1-score

6. Conclusion

According to the data taken coursera and by using WEATHER, ROADCOND, LIGHTCOND, ADDRTYPE, JUNCTIONTYPE as predictors we find that we can't rely very much on them since the best prediction percent is less than 80% and even if we use more predictors from the data we will find that these predictors have a lot of biased information like speeding for example in this case we don't know the type of drivers that are using the road and if they are speeders or not

Thanks for reading