4/8/2020

CST8390_012 Final Project

Decision Trees- Motor Vehicle Collisions – Crashes & NYC Weather

Rezansoff Karl
040955782

Li Min
040930563

a. How likely are people to get injured in a motor vehicle collision in Bronx Borough of New York City?

b. What factors make people more likely to be injured?

Table of Contents

Introduction

This our Business Intelligence Final Project, we are using J -48 Decision Tree to analyze a combined dataset which includes Motor Vehicle Collisions – Crashes and NYC Weather.

Step 1. Data Collection

We combined two datasets to do our analyzing.

The first one is:

Motor Vehicle Collisions - Crashes

Retrieved from:

https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/data

This dataset was found on the New York City Open Data website. The Motor Vehicle Collisions crash dataset contains details on crash events with each row representing a crash event. The Motor Vehicle Collisions dataset contains information from all police reported motor vehicle collisions in NYC.

The second one is:

NYC Weather

Retrieved from:

https://www.ncdc.noaa.gov/

This dataset was found on the National Centers for Environmental Information website. The NYC Weather Dataset contains details on weather records with each row representing each day's weather record.

Motor Vehicle Collisions - Crashes Dataset Attributes:

Crash Date	On which day a crash happened

Crash Time	When did a crash happen in a day
Borough	Which borough had a collision
Zip Code	Zip code of area where collision happened
Latitude	Latitude of collision happened area
Longitude	Longitude of collision happened area
Location	Location of crash
On Street Name	On which street the collision happened
Cross Street Name	Cross which street the collision happened
Off Street Name	Off which street the collision happened
Number of Persons Injured	Number of persons got injured in one collision
Number of Persons Killed	Number of persons got killed in one collision
Number of Pedestrians Injured	Number of pedestrians got injured in one collision
Number of Pedestrians Killed	Number of pedestrians got killed in one collision
Number of Cyclist Injured	Number of cyclists got injured in one collision
Number of Cyclist Killed	Number of cyclists got killed in one collision
Number of Motorist Injured	Number of motorists got injured in one collision
Number of Motorist Killed	Number of motorists got killed in one collision
Contributing Factor Vehicle1	Reasons made the crash happen like Aggressive Driving/Road Rage, Alcohol Involvement, Animals Action, Backing Unsafely, Brakes Defective, Cell Phone(hand-Held), Cell Phone(hands-free), etc.
Contributing Factor Vehicle2	Same reasons as Contributing Factor Vehicle1, but for the second vehicle in the crash

Contributing Factor Vehicle3	Same reasons as Contributing Factor Vehicle1, but for the third vehicle in the crash
Contributing Factor Vehicle4	Same reasons as Contributing Factor Vehicle1, but for the fourth vehicle in the crash
Contributing Factor Vehicle5	Same reasons as Contributing Factor Vehicle1, but for the fifth vehicle in the crash
Collision ID	Id that was given to collisions when reported
Vehicle Type Code1	Vehicle Type like: Bicycle, Truck, Car, Motorbike, School bus, Ambulance, etc.
Vehicle Type Code2	Same description as VehicleTypeCode1, but for the second vehicle in the accident.
Vehicle Type Code3	Same description as VehicleTypeCode1, but for the third vehicle in the accident.
Vehicle Type Code4	Same description as VehicleTypeCode1, but for the fourth vehicle in the accident.
Vehicle Type Code5	Same description as VehicleTypeCode1, but for the fifth vehicle in the accident.

NYC Weather Dataset Attributes:

Station	Unique identifier for the weather station (dataset is from only one weather station).
Name	Name of the weather station. All the data is from Farmingdale Republic Airport, NY
Date	Represents date

PRCP	Amount of precipitation for the day in inches.
TMAX	The daily high temperature.
TMIN	The daily minimum temperature.

Step 2. Initial Predictions

a. We predict that during commuting time like morning and afternoon there would be more collisions.

b. We predict on weekdays there would be more crashes than weekends since many people need to go to work.

c. We predict weather, low temperature (winter collisions) and crash time are more likely to make people injured in a vehicle accident, such as driving in the rain during the evening.

To answer our question, we will construct a J48 Decision Tree and the rules that are output should confirm if our prediction is correct and give us insight on what factors make someone more likely to be injured in a motor vehicle accident in Bronx borough of NYC.

Step 3. Preprocessing

Upon opening our dataset, we found that the 1.6M records did not fit in a single Excel sheet as the max is 1.04M records and we also found our laptops were unable to reliably run general excel tasks on a dataset this large without crashing. To cut down on the data we decided to only use data from the Bronx borough.

Upon reviewing the data, we noticed less records and missing values before 2015, so we cut out those values (22,560 records) and our model will use values from January 2016 – 2019. So we used data from the years 2016-2018 for our training set and will test our model on the 2019 data, which in total gives us 86,307 records.

Outliers we found were in the Number of Persons Injured Column (bus accidents), however we don't think it will skew our data as we converted that attribute to "someoneInjured" and only looking at if at least one person was injured.

To answer our initial questions, the factors we find relevant and would like to include in our model are the day of the week, time of day, vehicle type, precipitation, and temperature.

Attributes needed to be converted:

a. Converted NUMBER OF PERSONS INJURED into someoneInjured(0,1) with 1 indicating at least 1 person was injured in the incident.
b. Converted CRASH DATE into a nominal value IsWeekDay(0,1) with 1 indicating date of the crash was between Monday - Friday.
c. Converted CRASH TIME into nominal value CrashTime:
    - IF time < 12:00pm, it's morning,
    - IF time < 18:00pm, it's afternoon,
    - Else, it's evening.
d. Converted VEHICLE TYPE CODE 1 from more than 100 unique instances to "Car, Work, Taxi, Bike/Motorcycle, and Other/NS". All null or empty values were put into the Other/NS category.
e. Converted PRCP to isPRCP(0,1) with 1 indicating there was precipitation that day.
f. Combined TMAX and TMIN into TAVG by averaging the two (TMAX + TMIN) / 2, converting to Celsius and then created categories:
    - <-10
    - -10 – 0
    - 1 – 10
    - 11 – 21
    - >21

Lastly, we found missing values in the Vehicle Type column which we classified as Other/NS and we also noticed two days were missing from our weather dataset, so we keyed in the precipitation and temperature data manually after searching for the weather for those days.

Attributes Selected for our train and test datasets:

SomeoneInjured	Nominal value (0,1) 0 stands for not injured, 1 stands for injured. This column is used to indicate if someone was injured in the accident.
IsWeekday	Nominal value (0,1) 0 stands for weekdays, 1 stands for weekends. Instead of using the date of the accident, we computed the day of the week and created a nominal value to indicate if it occurred on a weekday or weekend. We predict that if it's weekdays, there would be more accidents. If it's weekends, there would be less collisions.
CrashTime	Nominal value (Morning, afternoon, evening) Instead of using the time of the accident we categorized this attribute as (Morning, Afternoon, Evening). We think if it's on commuting time that would influence the total amount of collisions, if there is during the night, there wouldn't be so many accidents.
VehicleTypeCode1	Nominal value for type of vehicle (Car, Work, Taxi, Bike/Motorcycle, and Other/NS).
isPRCP	Nominal value(0,1) with 1 indicating precipitation that day, and 0 for no precipitation.
TAVG	Nominal value for average temperature (<-10, -10–0, 1-10, 11-21, and >21)

Bronx_Train_Processed: collision records from 2016 to 2018.

Bronx_Train_Processed.arff

```
@relation Bronx_Train_Processed-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

@attribute somoneInjured {0,1}
@attribute isWeekday {0,1}
@attribute 'CRASH TIME' {Afternoon,Morning,Evening}
@attribute 'VEHICLE TYPE CODE 1' {'Passenger Vehicle',Other/NS,Work,Taxi,Bike/Motorcycle}
@attribute isPRCP {0,1}
@attribute TAVG {1-10,-10-0,<-10,11-21,>21}

@data
0,1,Afternoon,'Passenger Vehicle',0,1-10
0,1,Morning,'Passenger Vehicle',0,1-10
1,1,Evening,Other/NS,0,1-10
0,1,Morning,'Passenger Vehicle',0,1-10
0,1,Morning,'Passenger Vehicle',0,1-10
0,1,Morning,'Passenger Vehicle',0,1-10
1,1,Morning,'Passenger Vehicle',0,1-10
1,1,Evening,Other/NS,0,1-10
0,1,Afternoon,Other/NS,0,1-10
0,1,Afternoon,'Passenger Vehicle',0,1-10
0,1,Morning,'Passenger Vehicle',0,1-10
0,1,Morning,'Passenger Vehicle',0,1-10
0,1,Evening,'Passenger Vehicle',0,1-10
0,1,Morning,'Passenger Vehicle',0,1-10
```

Bronx_Test_Processed: collision records of 2019.

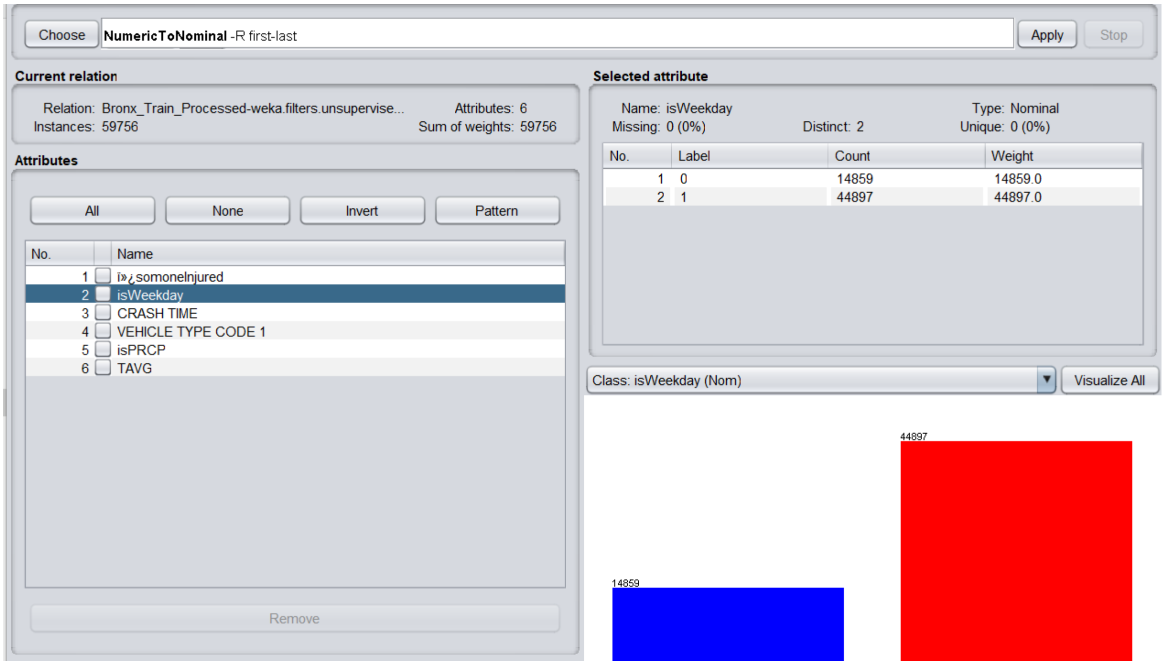
Bronx_Test_Processed.arff

```
@relation Bronx_Test_Processed-weka.filters.unsupervised.attribute.StringToNominal-R1-weka.filters.unsupervised.attribute.NumericToNominal-R2-weka.filters.unsupervised

@attribute somoneInjured {0,1}
@attribute isWeekday {0,1}
@attribute 'CRASH TIME' {Afternoon,Morning,Evening}
@attribute 'VEHICLE TYPE CODE 1' {'Passenger Vehicle',Other/NS,Work,Taxi,Bike/Motorcycle}
@attribute isPRCP {0,1}
@attribute TAVG {1-10,-10-0,<-10,11-21,>21}

@data
?,1,Morning,'Passenger Vehicle',1,1-10
?,1,Morning,'Passenger Vehicle',1,1-10
?,1,Afternoon,'Passenger Vehicle',1,1-10
?,1,Morning,'Passenger Vehicle',1,1-10
?,1,Evening,'Passenger Vehicle',1,1-10
?,1,Afternoon,'Passenger Vehicle',1,1-10
?,1,Morning,'Passenger Vehicle',1,1-10
?,1,Evening,'Passenger Vehicle',1,1-10
?,1,Afternoon,'Passenger Vehicle',1,1-10
?,1,Morning,Taxi,1,1-10
?,1,Morning,'Passenger Vehicle',1,1-10
```

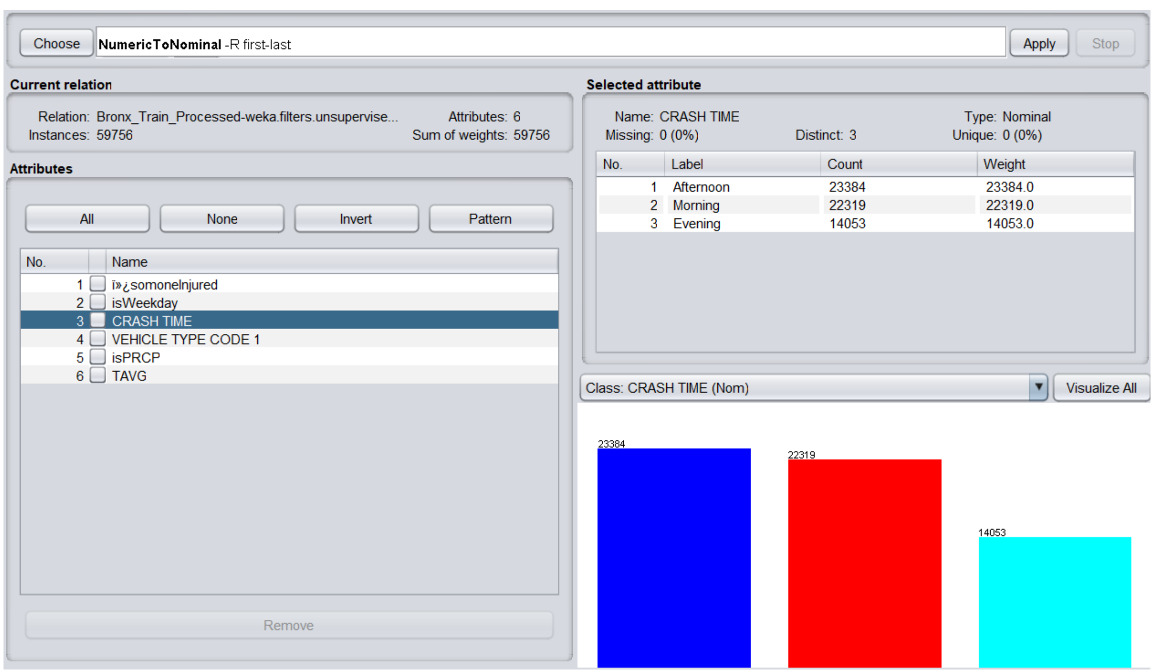
Step 4. Data Analysis

Upon viewing the data we found:

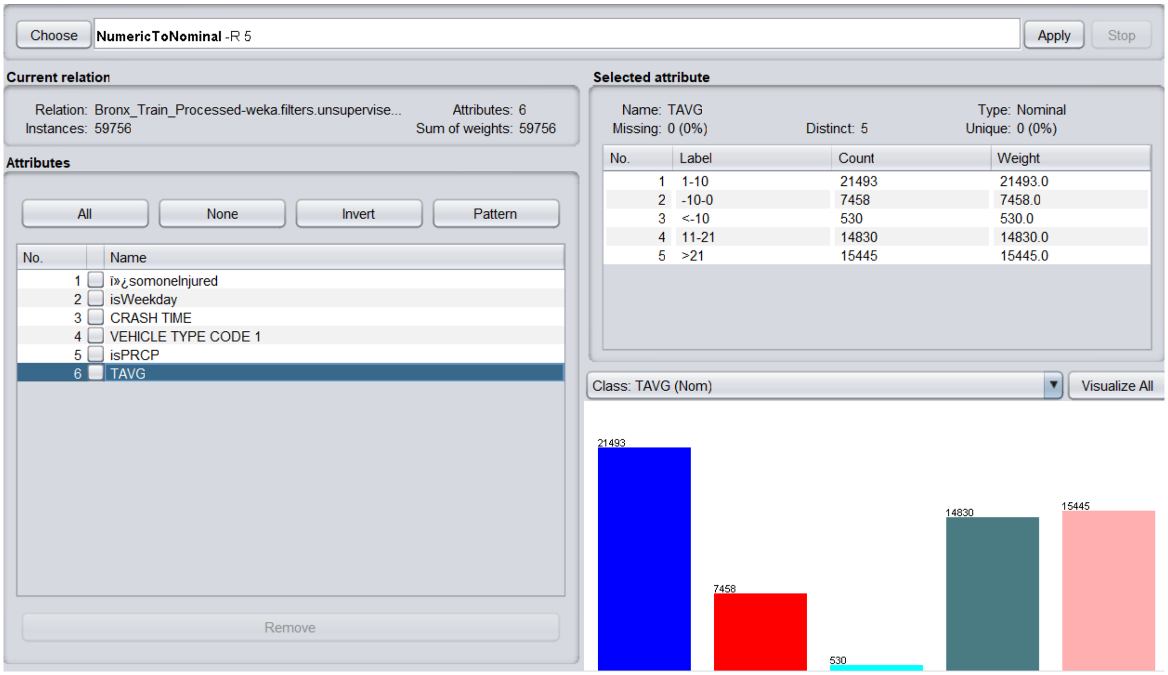
a. Most accidents occurred on weekdays



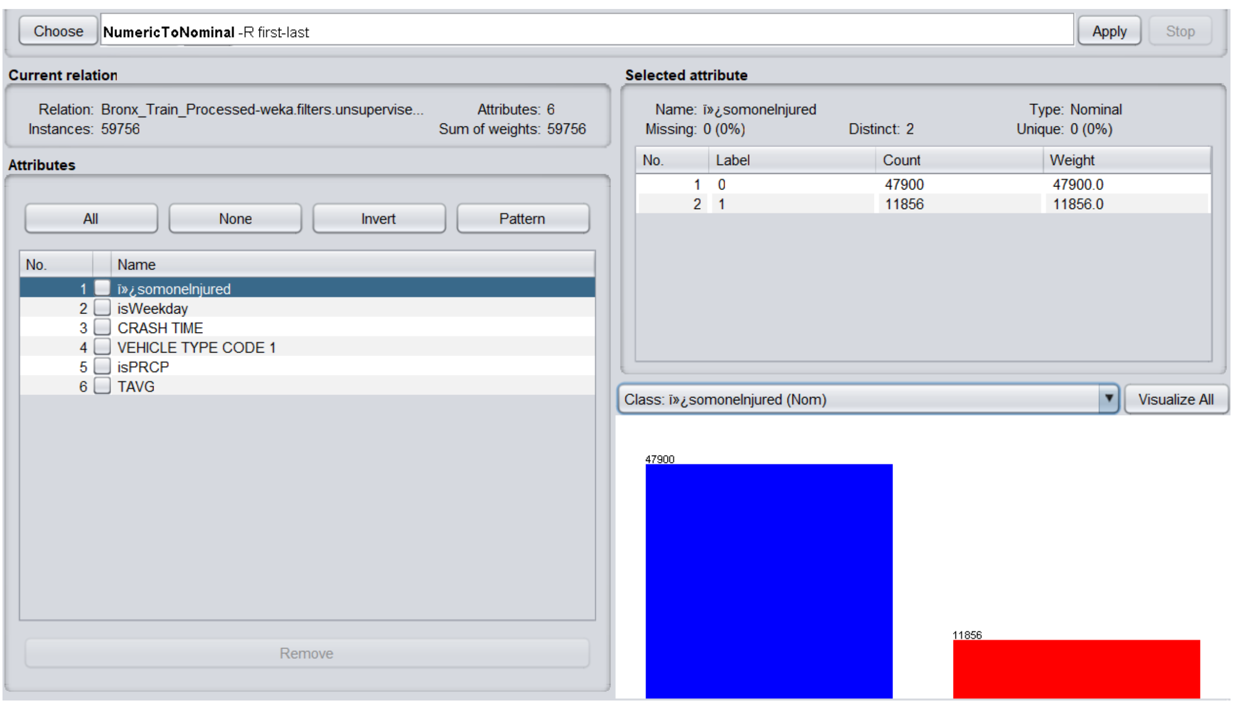
b. Most accidents are in the morning and afternoon



c. Most accidents occurred in weather between 1-10 degrees Celsius.



d. Injuries occur in 19.8% of the motor vehicle accidents (11856 collisions with injuries out of the 59756 collisions in total)



Step 5. Results & Summary

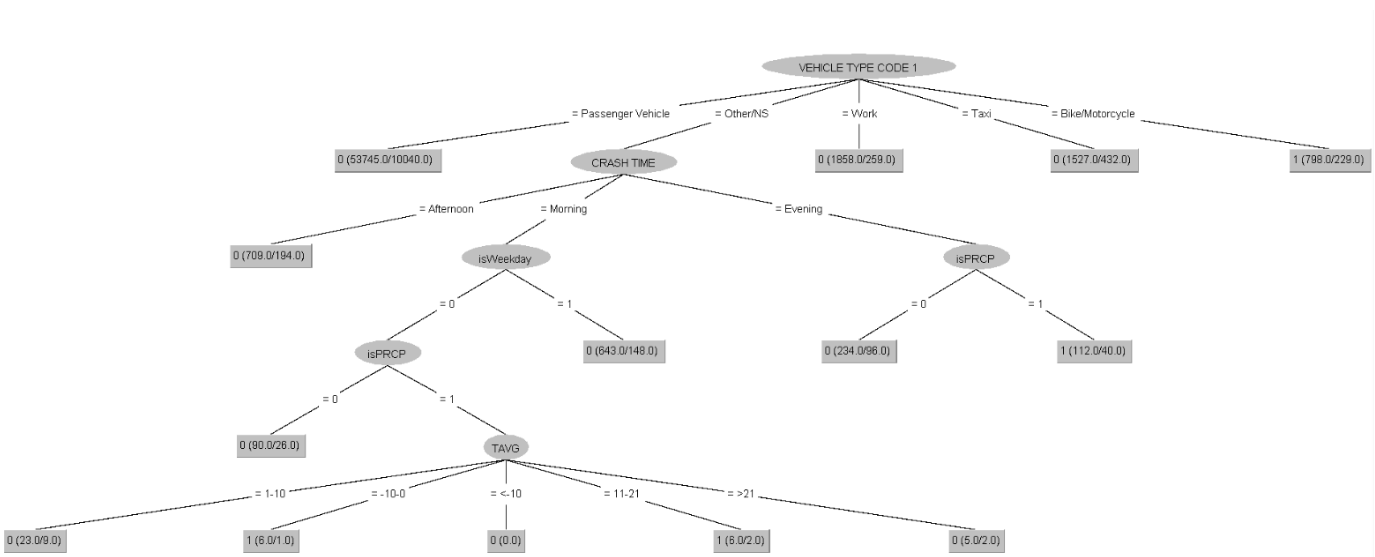
Results from J48 tree for our test dataset using default values and a confidence factor of 0.4:

```
J48 pruned tree
------------------

VEHICLE TYPE CODE 1 = Passenger Vehicle: 0 (53745.0/10040.0)
VEHICLE TYPE CODE 1 = Other/NS
|   CRASH TIME = Afternoon: 0 (709.0/194.0)
|   CRASH TIME = Morning
|   |   isWeekday = 0
|   |   |   isPRCP = 0: 0 (90.0/26.0)
|   |   |   isPRCP = 1
|   |   |   |   TAVG = 1-10: 0 (23.0/9.0)
|   |   |   |   TAVG = -10-0: 1 (6.0/1.0)
|   |   |   |   TAVG = <-10: 0 (0.0)
|   |   |   |   TAVG = 11-21: 1 (6.0/2.0)
|   |   |   |   TAVG = >21: 0 (5.0/2.0)
|   |   isWeekday = 1: 0 (643.0/148.0)
|   CRASH TIME = Evening
|   |   isPRCP = 0: 0 (234.0/96.0)
|   |   isPRCP = 1: 1 (112.0/40.0)
VEHICLE TYPE CODE 1 = Work: 0 (1858.0/259.0)
VEHICLE TYPE CODE 1 = Taxi: 0 (1527.0/432.0)
VEHICLE TYPE CODE 1 = Bike/Motorcycle: 1 (798.0/229.0)

Number of Leaves  :     14

Size of the tree :      20
```

Summary and Confusion Matrix

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       48251               80.7467 %
Incorrectly Classified Instances     11505               19.2533 %
Kappa statistic                          0.0742
Mean absolute error                      0.309
Root mean squared error                  0.3932
Relative absolute error                 97.1358 %
Root relative squared error             98.6076 %
Total Number of Instances            59756

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.994    0.945    0.809      0.994    0.892      0.155    0.541     0.817     0
                 0.055    0.006    0.685      0.055    0.101      0.155    0.541     0.259     1
Weighted Avg.    0.807    0.759    0.785      0.807    0.735      0.155    0.541     0.706

=== Confusion Matrix ===

     a     b   <-- classified as
 47602   298 |     a = 0
 11207   649 |     b = 1
```

Decision Tree



Step 6. After Using Test Dataset

The following is the result:

1	prediction margin	predicted ?somoneInjured	somoneInjured	isWeekda	CRASH	VEHICL	isPRCP	TAVG
21	-0.426065	1	?	1	Afternoon	Bike/Motc	1	10-Jan
91	-0.426065	1	?	1	Morning	Bike/Motc	0	10-Jan
361	-0.426065	1	?	1	Morning	Bike/Motc	0	-10
446	-0.285714	1	?	1	Evening	Other/NS	1	10-Jan
465	-0.426065	1	?	1	Morning	Bike/Motc	0	-10
575	-0.426065	1	?	1	Morning	Bike/Motc	0	-10
600	-0.426065	1	?	1	Afternoon	Bike/Motc	0	-10
720	-0.426065	1	?	1	Afternoon	Bike/Motc	0	-10
2992	-0.426065	1	?	1	Evening	Bike/Motc	1	10-Jan
3014	-0.426065	1	?	1	Afternoon	Bike/Motc	1	10-Jan
3129	-0.426065	1	?	0	Afternoon	Bike/Motc	1	10-Jan
3130	-0.285714	1	?	0	Evening	Other/NS	1	10-Jan

results

323 of 22431 records found

1	isWeekday	CRASH TIM	CRASH TIM	NUMBER OF PERSONS INJURED	somoneInjured	VEHICLE TYPE CODE 1	PRCP	isPRCP	TAVG	TAVG
59760	1	12:58:00	Afternoon	1	1	Passenger Vehicle	0.1	1	9.722222	1-10
59765	1	20:17:00	Evening	3	1	Passenger Vehicle	0.1	1	9.722222	1-10
59768	1	01:15:00	Morning	1	1	Passenger Vehicle	0.1	1	9.722222	1-10
59777	1	16:00:00	Afternoon	1	1	Bike/Motorcycle	0.1	1	9.722222	1-10
59778	1	03:22:00	Morning	1	1	Passenger Vehicle	0.1	1	9.722222	1-10
59794	1	10:10:00	Morning	1	1	Other/NS	0.1	1	9.722222	1-10
59803	1	12:00:00	Afternoon	1	1	Passenger Vehicle	0.1	1	9.722222	1-10
59806	1	00:00:00	Morning	1	1	Passenger Vehicle	0	0	1.944444	1-10
59813	1	15:45:00	Afternoon	1	1	Taxi	0	0	1.944444	1-10
59903	1	19:00:00	Evening	1	1	Passenger Vehicle	0	0	3.333333	1-10
59906	1	17:21:00	Afternoon	1	1	Passenger Vehicle	0	0	3.333333	1-10
59913	1	14:48:00	Afternoon	1	1	Passenger Vehicle	0	0	2.222222	1-10
59914	1	19:30:00	Evening	1	1	Passenger Vehicle	0	0	2.222222	1-10

Sheet1 WeatherData

4778 of 86307 records found

The Bronx Borough had records on 22,431 motor vehicle accidents and our model predicted 323 or 1.03% of those crashes had injuries, while the actual number of crashes with injuries was 4778 or 5.5%.

Vehicle Type was the root node and predicted:

- All passenger vehicle drivers to not be injured.
- All bike/motorcycle drivers to be injured.
- All work vehicle drivers to not be injured.

Vehicle Type Other/NS did create additional leaves:

- More likely to be injured if its evening, and precipitation.
- More likely to be injured if its morning, precipitation, and between -10-0 or 11-21 degrees Celsius.
- All afternoon crashes classified as not injured.

Step 7. Comparing Our Prediction & Results

- As we predicted most accidents occur during commute time (morning and afternoon), our model shows most of collisions happened in the morning and afternoon which means our prediction is correct, however more injuries occur in morning and evening crashes.
- We predicted that most crashes happen on weekdays, our model shows the same result, so this prediction is correct, too.
- Precipitation does make people more likely to be injured in an accident, however we were expecting precipitation and low temperatures (winter collisions) to be a bigger factor.

Conclusion & Recommendation

In conclusion, crashes occur in 19.8% of motor vehicle accidents and driving a bike or motorcycle is the biggest factor in predicting if people will be injured in a motor vehicle accident. In addition, precipitation, and driving in the morning or evening also make people more likely to be injured if people were to be in a motor vehicle collision.

Moving forward, to further improve our model we think other factors outside of what was included in our dataset should be considered, such as reason for collision, age of driver/passengers, and location.

We think more attributes could have been useful, as our tree only made additional rules for vehicle types classified as (Other/NS). The Vehicle Type column was very inconsistent, and we subjectively categorized the instances into 5 categories, so the errors may have skewed the results.