Prediction of Accident Severity

Contents

Prediction of Accident Severity

1. Introduction

1.1. Background

In traffic situations, passengers are prone to accidents on the roads. This can be due to different factors such as the weather conditions, the road conditions, the light conditions amongst other factors. These attributes are being stored by the traffic system in Seattle. It is highly recommended to be able to predict the severity of an accident based on the factors available to prepare for the casualty before the accident occurs.

1.2. Problem

The dataset provided for the Seattle city contains a total of 38 attributes (relating to the accidents that occur on the road) and the labelled data which describes the fatality of an incident. Given this dataset, the aim of this project is to select the necessary attributes that will be used to build a model that will help to predict the severity of an accident.

1.3. Interest

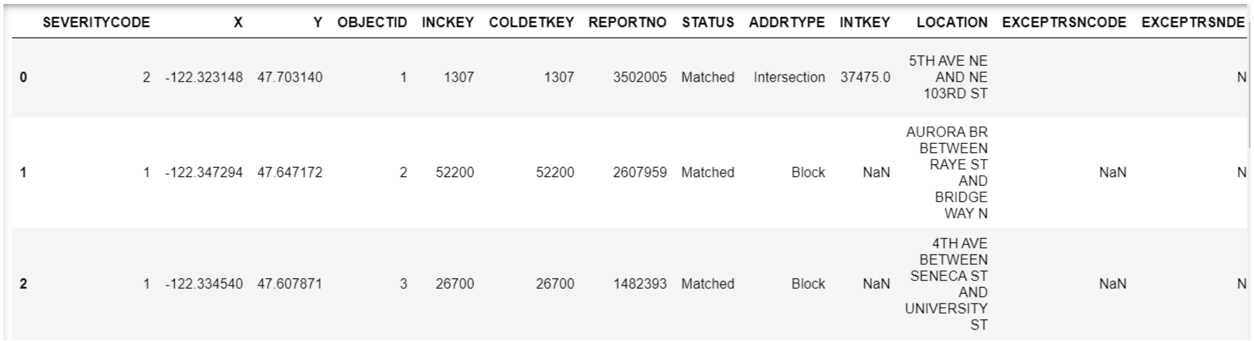
Residents of Seattle will find this helpful in predicting how severe an accident will be if they get into one based on the factors available. It will also be useful for traffic attendants and paramedic to prepare for accidents likely to happen. This will help reduce causality.

2. Data acquisition and cleaning

2.1. Data Sources

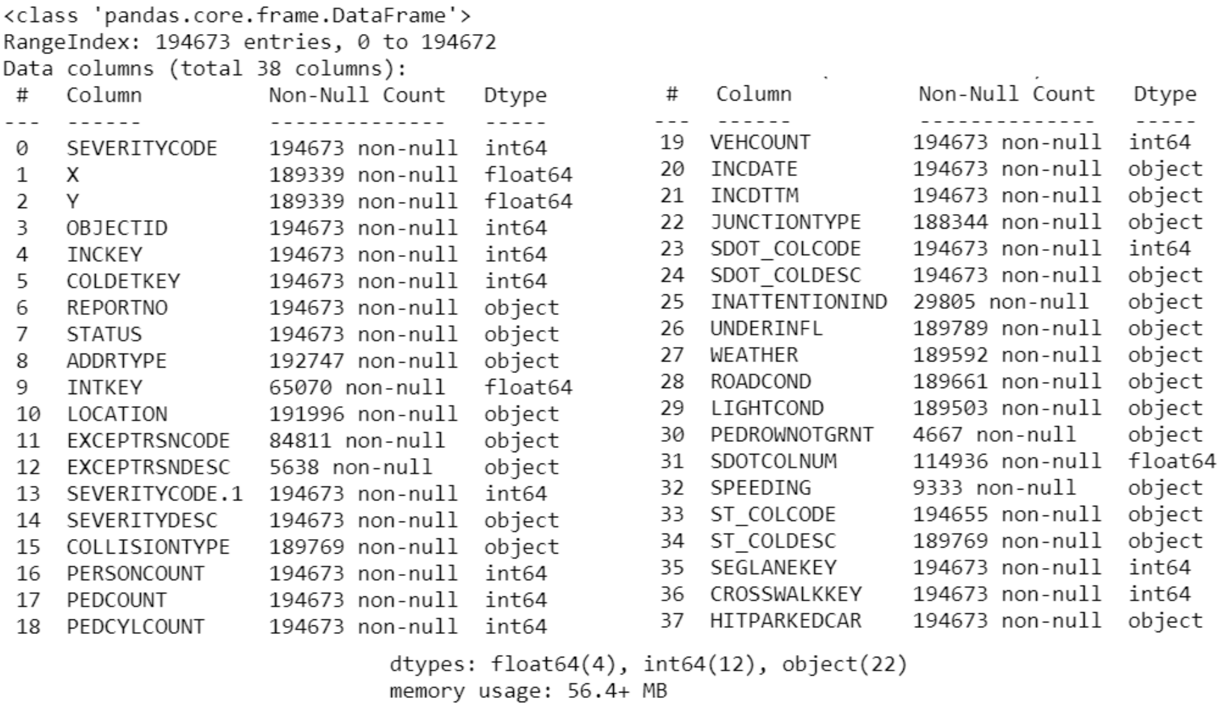
The dataset for Seattle road accidents was provided containing a total of 194673 observations and 37 attributes with many of them being categorical attributes.

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv



	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	LOCATION	EXCEPTRSNCODE	EXCEPTRSNDE
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	5TH AVE NE AND NE 103RD ST		N
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	AURORA BR BETWEEN RAYE ST AND BRIDGE WAY N	NaN	N
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	4TH AVE BETWEEN SENECA ST AND UNIVERSITY ST	NaN	N

Figure 1 Data frame from data read from given source.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 194673 entries, 0 to 194672
Data columns (total 38 columns):
 #   Column          Non-Null Count   Dtype          #   Column          Non-Null Count   Dtype
---  ------          --------------   -----         ---  ------          --------------   -----
 0   SEVERITYCODE    194673 non-null  int64          19  VEHCOUNT        194673 non-null  int64
 1   X               189339 non-null  float64        20  INCDATE         194673 non-null  object
 2   Y               189339 non-null  float64        21  INCDTTM         194673 non-null  object
 3   OBJECTID        194673 non-null  int64          22  JUNCTIONTYPE    188344 non-null  object
 4   INCKEY          194673 non-null  int64          23  SDOT_COLCODE    194673 non-null  int64
 5   COLDETKEY       194673 non-null  int64          24  SDOT_COLDESC    194673 non-null  object
 6   REPORTNO        194673 non-null  object         25  INATTENTIONIND  29805 non-null   object
 7   STATUS          194673 non-null  object         26  UNDERINFL       189789 non-null  object
 8   ADDRTYPE        192747 non-null  object         27  WEATHER         189592 non-null  object
 9   INTKEY          65070 non-null   float64        28  ROADCOND        189661 non-null  object
 10  LOCATION        191996 non-null  object         29  LIGHTCOND       189503 non-null  object
 11  EXCEPTRSNCODE   84811 non-null   object         30  PEDROWNOTGRNT   4667 non-null    object
 12  EXCEPTRSNDESC   5638 non-null    object         31  SDOTCOLNUM      114936 non-null  float64
 13  SEVERITYCODE.1  194673 non-null  int64          32  SPEEDING        9333 non-null    object
 14  SEVERITYDESC    194673 non-null  object         33  ST_COLCODE      194655 non-null  object
 15  COLLISIONTYPE   189769 non-null  object         34  ST_COLDESC      189769 non-null  object
 16  PERSONCOUNT     194673 non-null  int64          35  SEGLANEKEY      194673 non-null  int64
 17  PEDCOUNT        194673 non-null  int64          36  CROSSWALKKEY    194673 non-null  int64
 18  PEDCYLCOUNT     194673 non-null  int64          37  HITPARKEDCAR    194673 non-null  object
                              dtypes: float64(4), int64(12), object(22)
                              memory usage: 56.4+ MB
```

Figure 2 Data frame information about available columns

2.2. Data Cleaning and Preprocessing

Out of the numerous attributes, only selected attributes were used because they relate to the severity code based on their description in the metadata. The attributes are: 'ADDRTYPE', 'COLLISIONTYPE', 'JUNCTIONTYPE', 'WEATHER','ROADCOND','LIGHTCOND', 'PERSONCOUNT','PEDCOUNT', 'VEHCOUNT', 'HITPARKEDCAR'. The other attributes were dropped either because they do not relate to the target variable or because they have a lot of missing values.

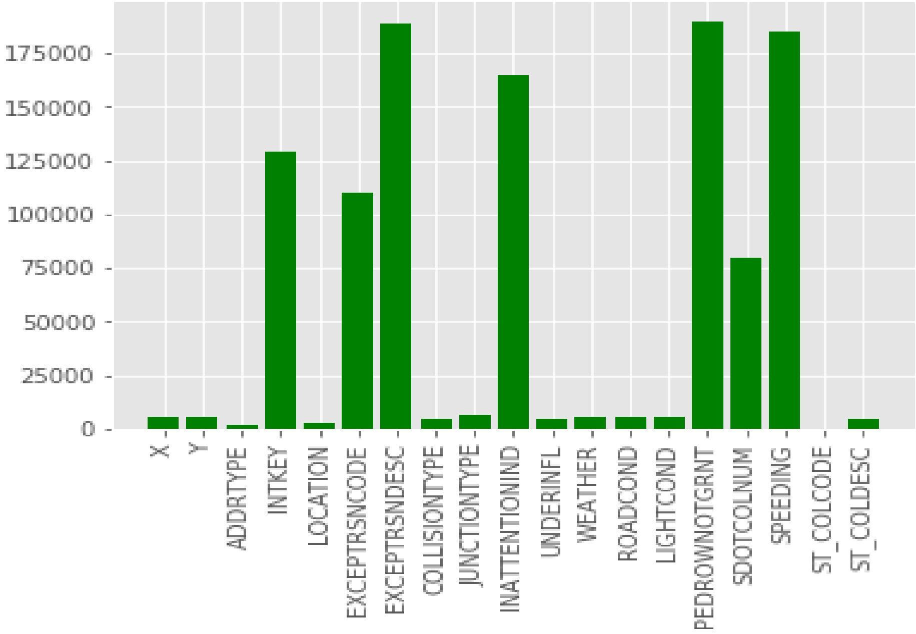


Figure 3 To analyze ratio of NaN across all selected columns

The following attributes are categorical values and needed to be changed to numerical values using the categorization of data.

- ADDRTYPE
- COLLISIONTYPE
- JUNCTIONTYPE
- UNDERINFL
- WEATHER
- ROADCOND
- LIGHTCOND
- SPEEDING
- HITPARKEDCAR

Applied required conditioning on columns like removal of rows with NaN values, type conversion etc.

3. Methodology

3.1. Exploratory Analysis

3.1.1 Relationship between ADDRTYPE and SEVERITYCODE

It was observed that 89% of the incidents that happened on the Alley has a severity code of 1 while the 11% has a severity code of 2. Also, majority of the incident that happened at the Block had a severity of 1. However, for incidents that occurred at intersection, half of the

incidents had severity code of 1 while the other half had the severity code of 2 as shown in the bar chart below.

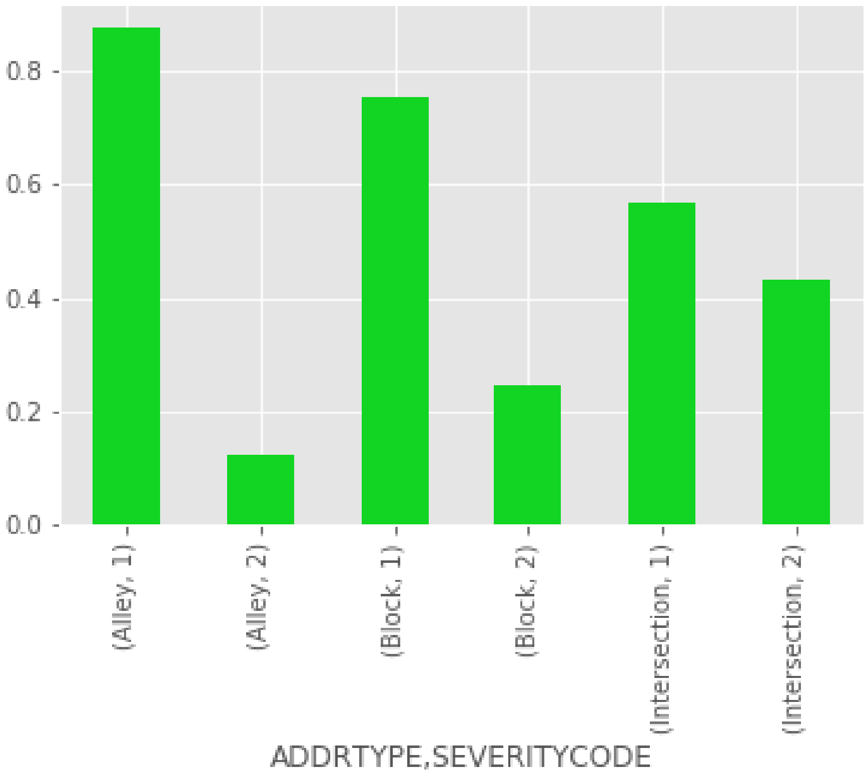


Figure 4 Address Type and Severity Code Relationship Analysis

3.1.2 Relationship between COLLISIONTYPE and SEVERITYCODE

It was observed that majority of incidents that involved Cycles or pedestrians had a severity of 2. However, for incidents in which the collision that involved a parked car and sideswipe, the severity code was 1.

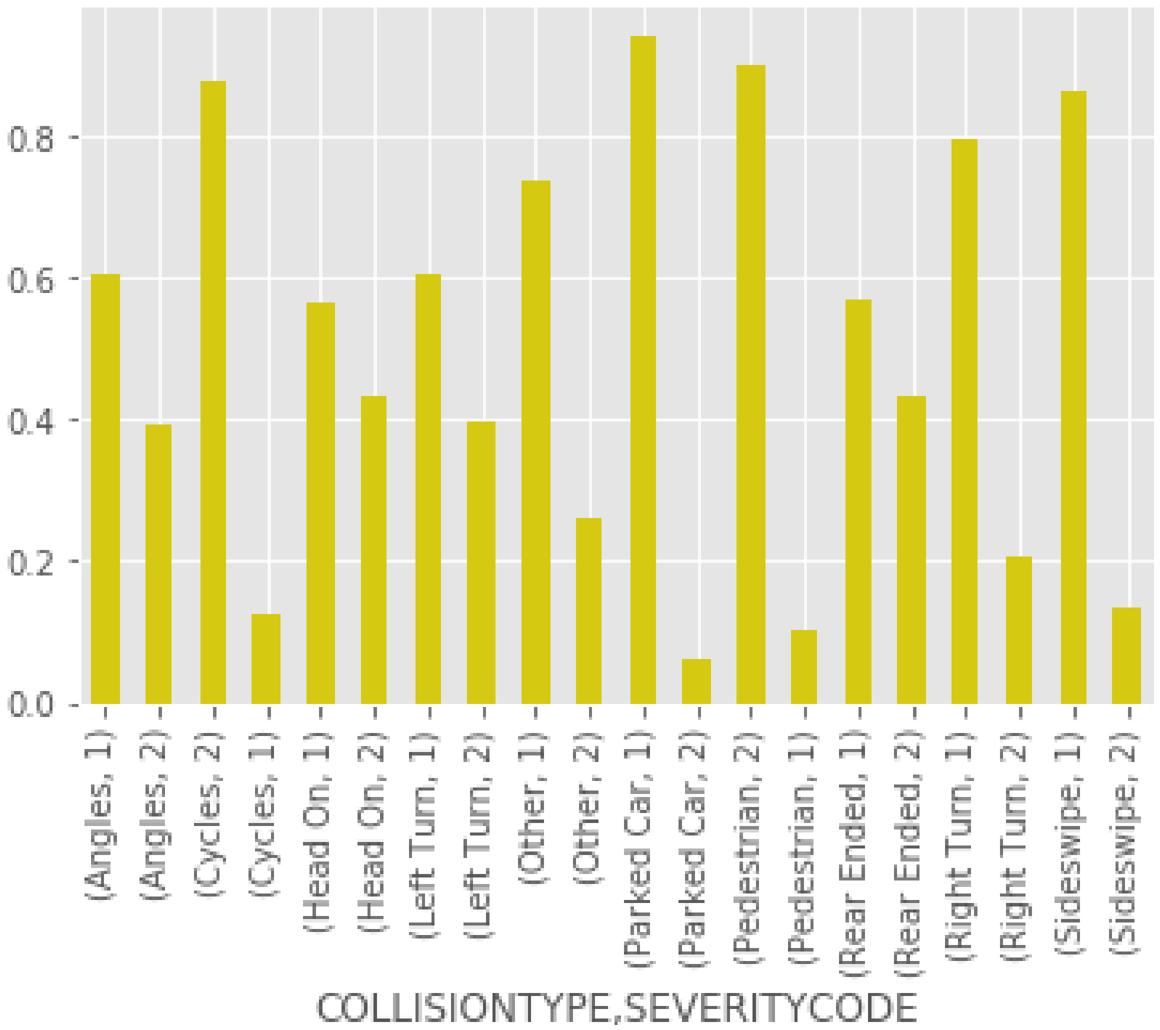


Figure 5 Collision Type and Severity Code Relationship Analysis

3.1.3 Relationship between VEHCOUNT and SEVERITYCODE

It is observed that VEHCOUNT is not directly related severity of accident, as we can see, there are accidents with severity of 2 and vehicle count is 0, whereas there are cases where vehicle count is 12, but accident of severity 1.

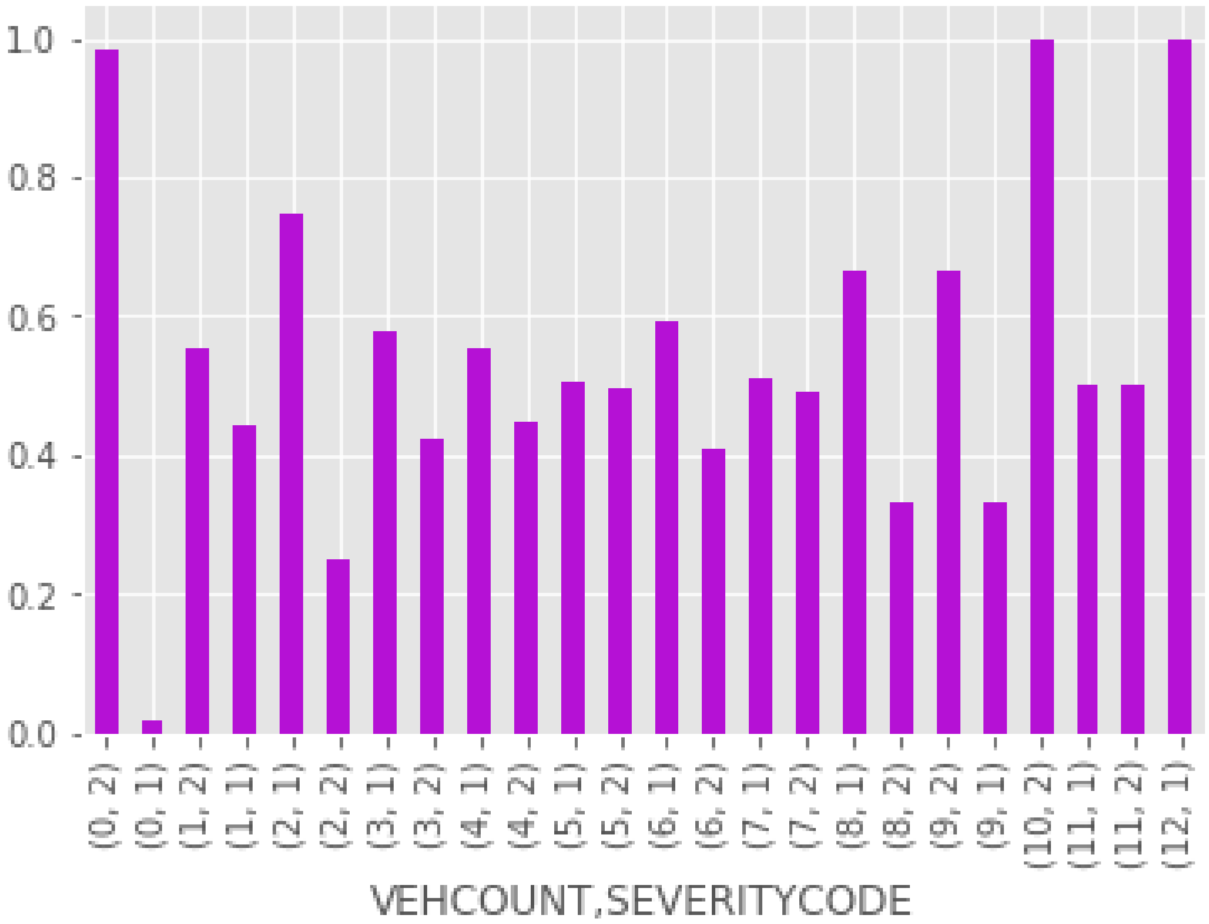


Figure 6 Vehicle Count and Severity Code Relationship Analysis

3.1.4 Relationship between JUNCTIONTYPE and SEVERITYCODE

Here observation is that JUNCTIONTYPE contributes in accident severity. JUNCTIONTYPE Mid-Block (not related to intersection) and At Intersection (but not related to intersection) contributed in more than 70% of accidents of severity 1. JUNCTIONTYPE At Intersection (intersection related), contributing 43% and Driveway Junction in 30% of severity 2 accidents. So JUNCTIONTYPE is going to be interesting feature for model.

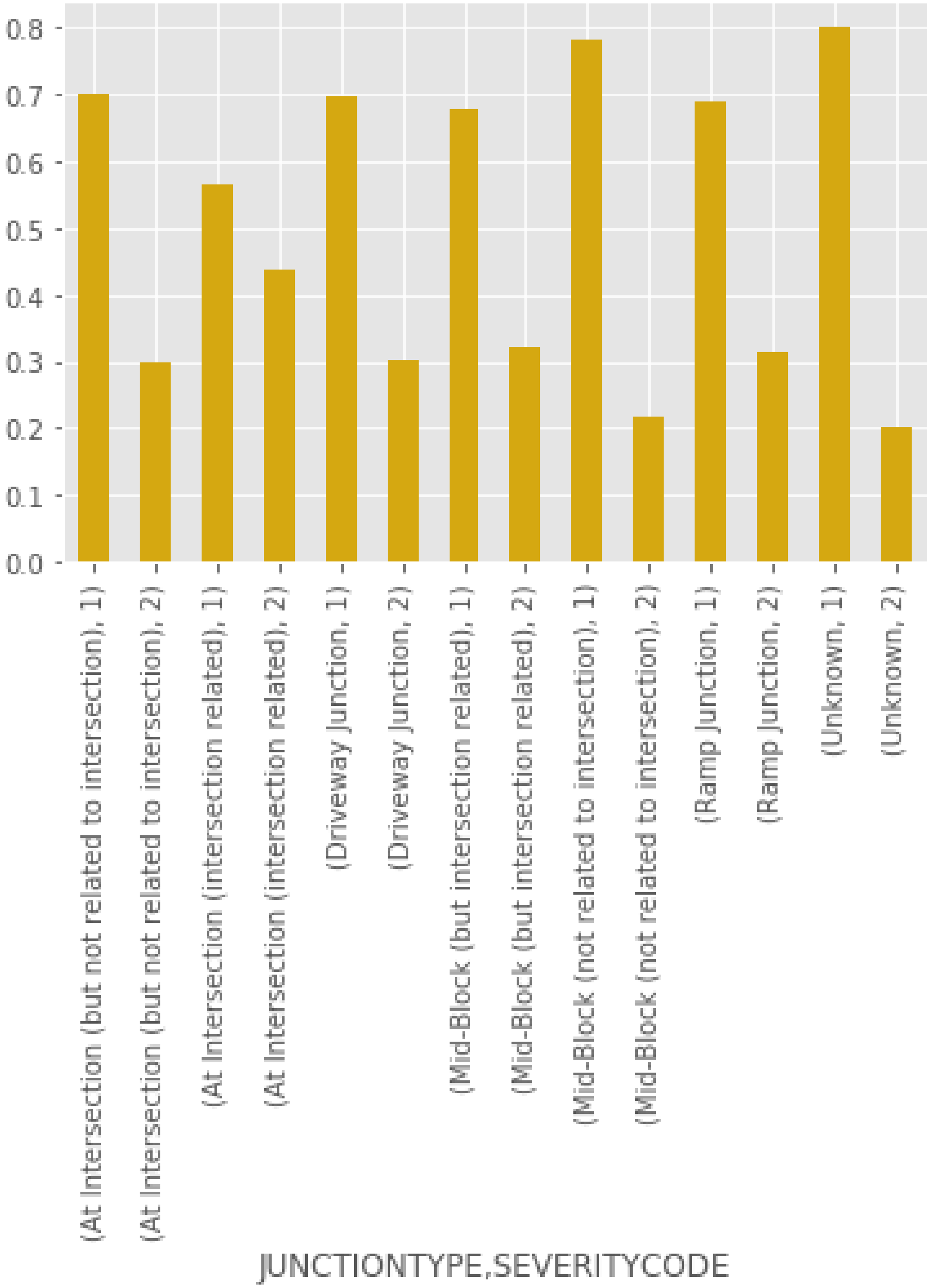


Figure 7 Junction Type and Severity Code Relationship Analysis

3.1.5 Relationship between UNDERINFL and SEVERITYCODE

Here relationship of accident severity and under influence of alcohol or drugs can be seen. We see almost similar percentage of severities in case of under influence and without it. So it must be seen in combination of other features like JUNCTIONTYPE, ROADCOND, LIGHTCOND.

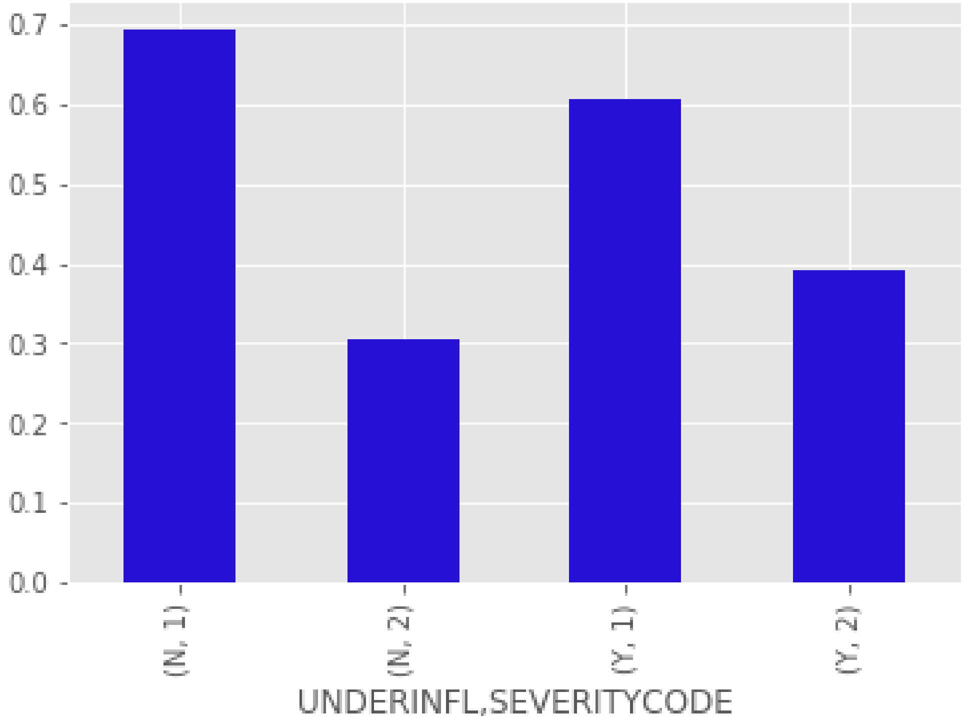


Figure 8 Under Influence of Alcohol or drugs and Severity Code Relationship Analysis

3.1.6 Relationship between WEATHER and SEVERITYCODE

It was observed that when the weather was Snowing or freezing rain, majority of the incident that occurred had a severity code of 1. However, most of the incidents that occurred when the weather was partly cloudy had a severity code of 2.

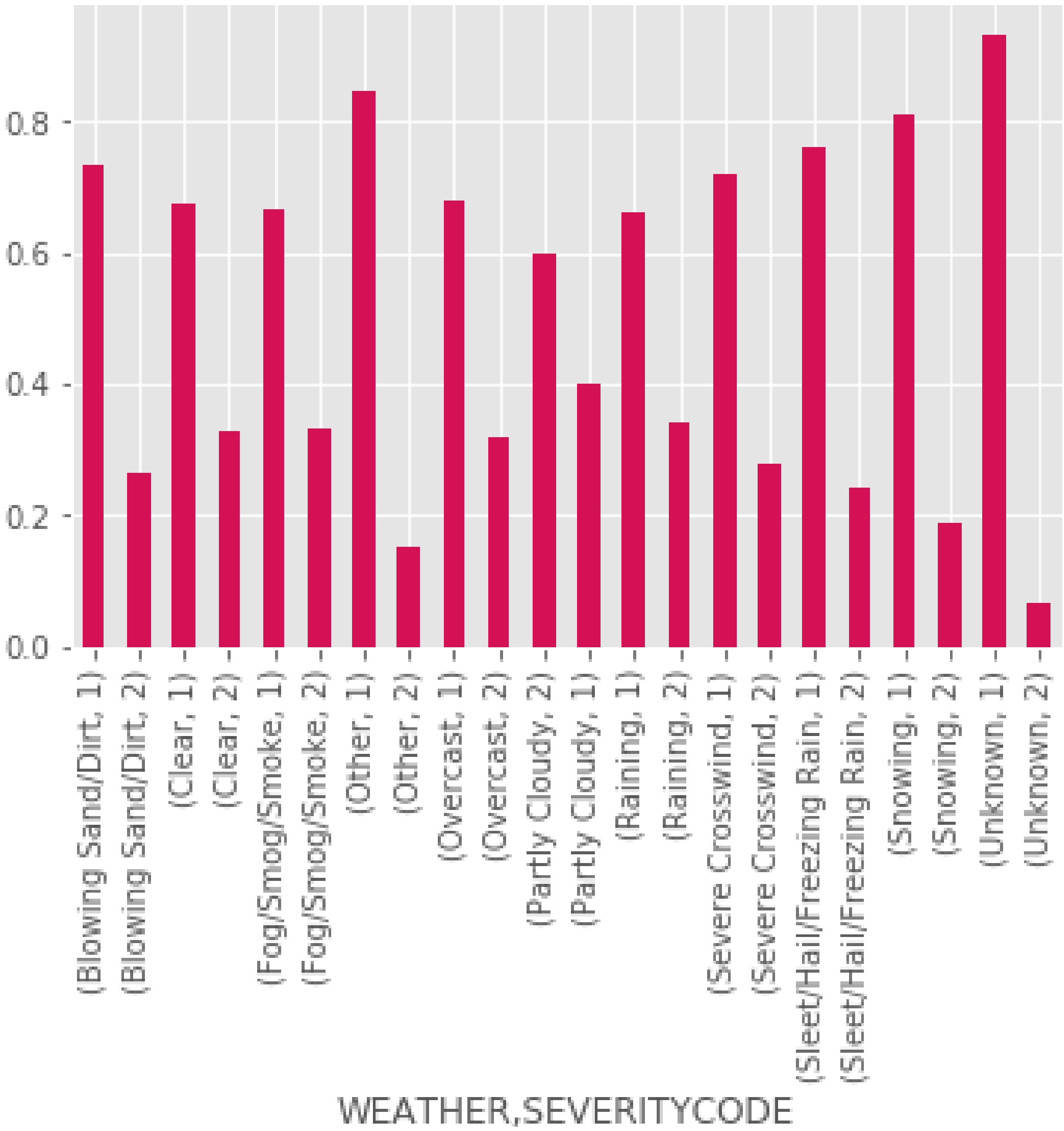


Figure 9 Weather Condition and Severity Code Relationship Analysis

3.1.7 Relationship between ROADCOND and SEVERITYCODE

It was observed that when the road condition was Ice or Snow/Slush, majority of the incident that occurred had a severity code of 1. However, most of the incidents that occurred when the road condition was Oil or wet, had a severity code of 2.

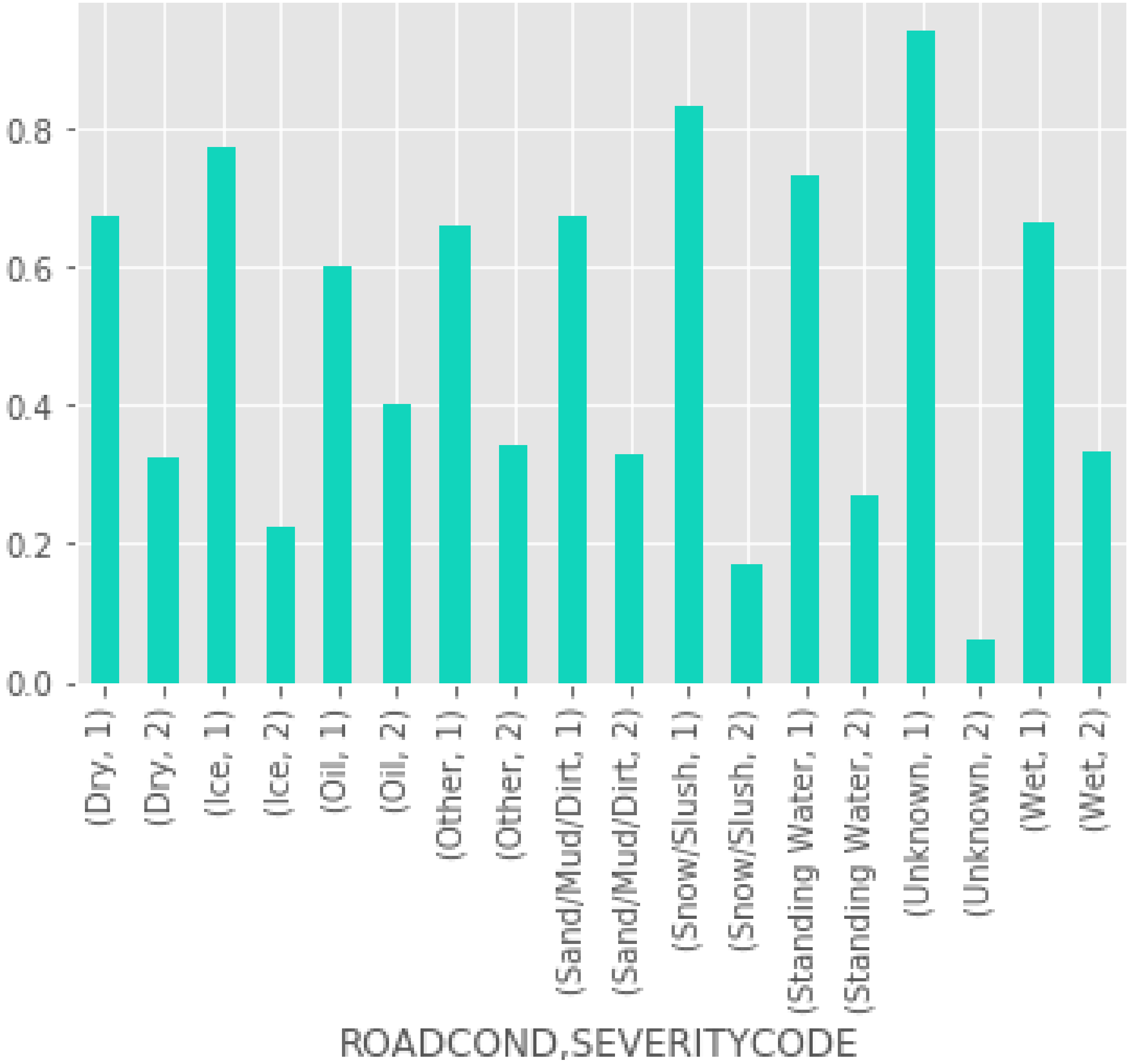


Figure 10 Road Condition and Severity Code Relationship Analysis

3.1.8 Relationship between LIGHCOND and SEVERITYCODE

It was observed that when the light condition was Dark - No Street Lights, Dark - Street Lights Off, Dark - Unknown Lighting, majority of the incident that occurred had a severity code of 1. However, almost 30% of the incidents that occurred with other light condition, had a severity code of 2. LIGHTCOND can be seen collectively with other features.

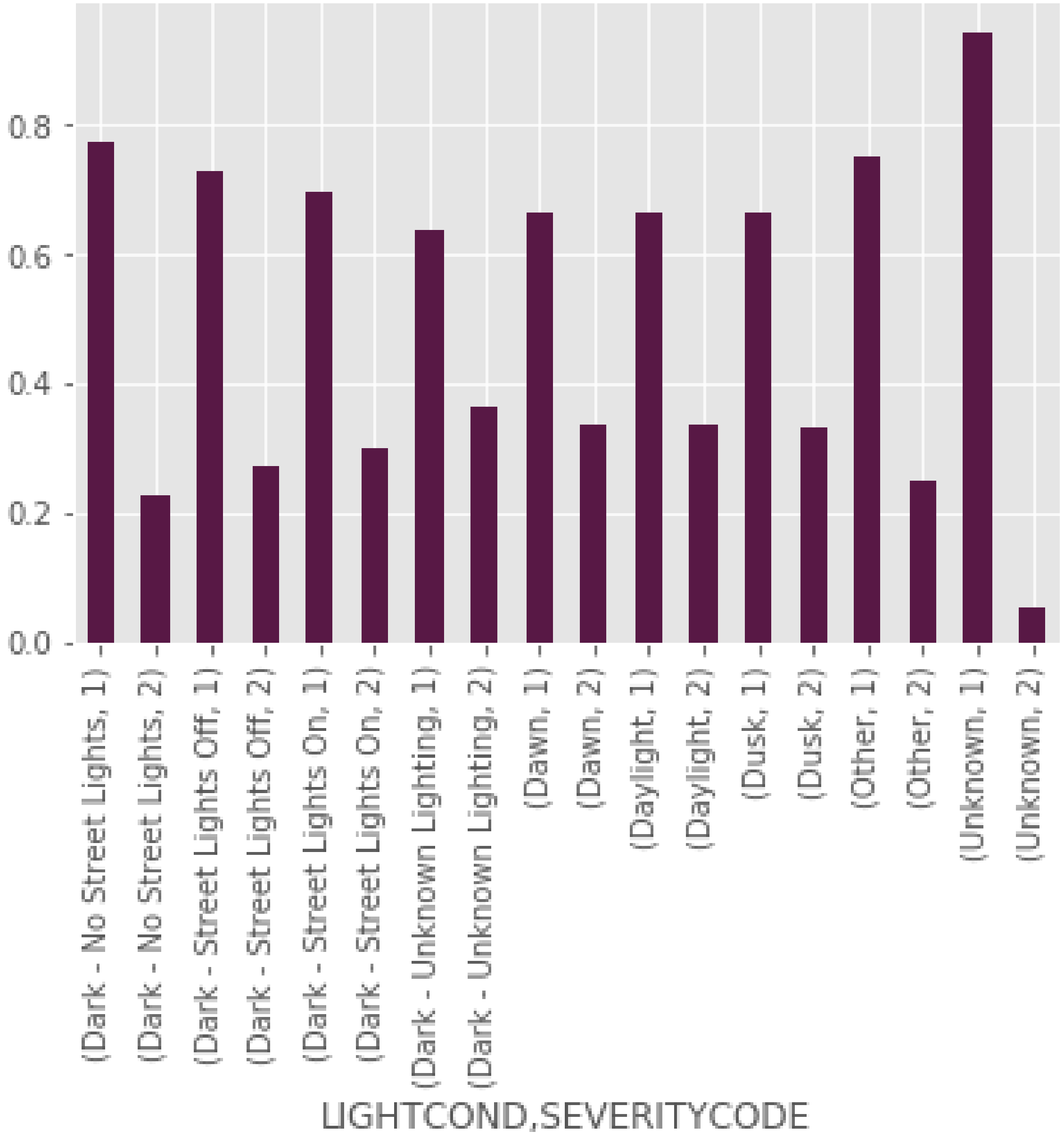


Figure 11 Light Condition and Severity Code Relationship Analysis

3.1.9 Relationship between SPEEDING and SEVERITYCODE

Here we can observe that relationship is a bit confusing, as car speeding and not speeding seen in almost 70% of accidents with severity 1. So better to combine it other features, to make model more robust.

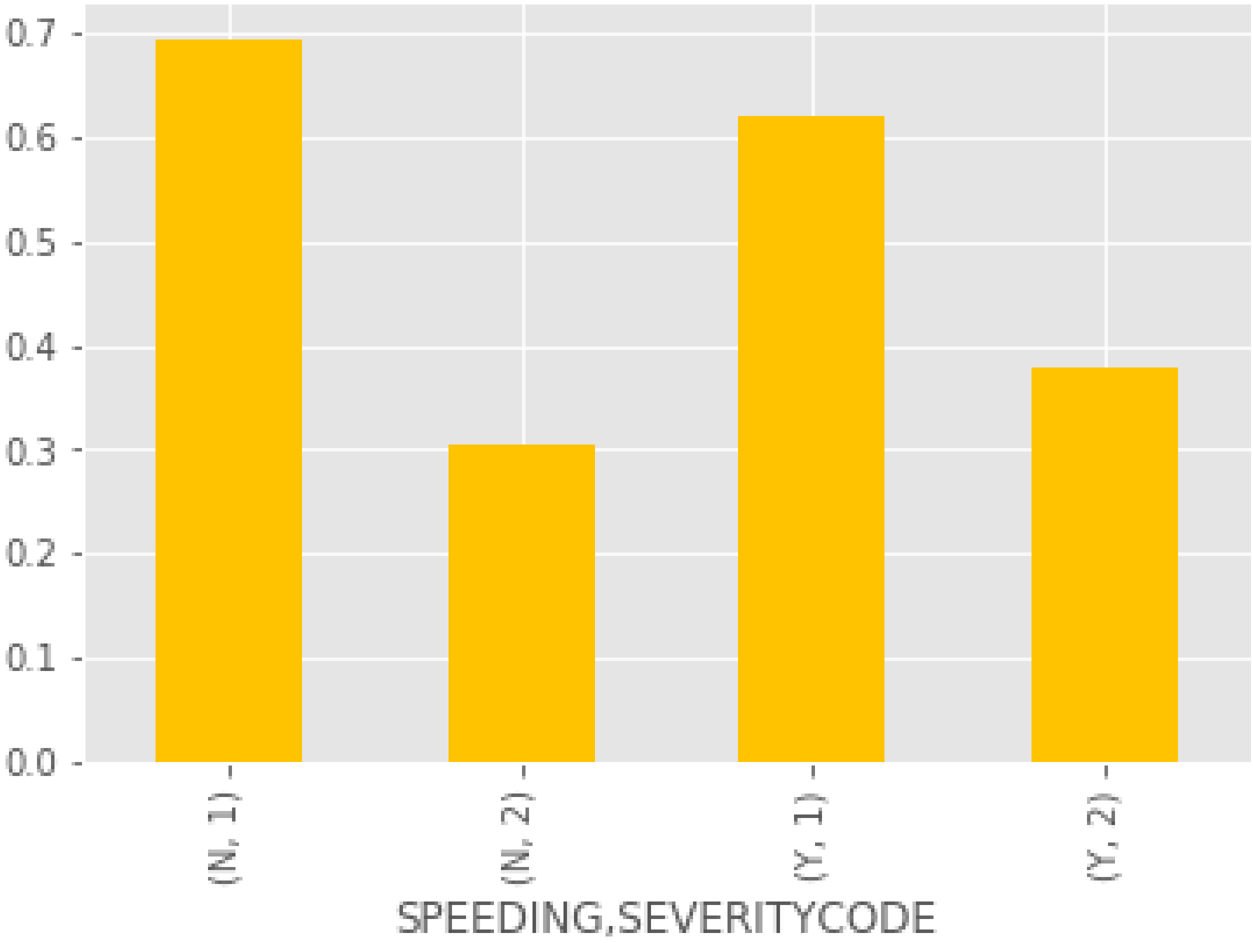


Figure 12 Car Speeding and Severity Code Relationship Analysis

3.1.10 Relationship between HITPARKEDCAR and SEVERITYCODE

We can observe that there are 92% accidents of severity 1, where parked car hit, whereas accidents severity of 2 can seen in lesser percentage, when parked car hit. It can be good feature for model.

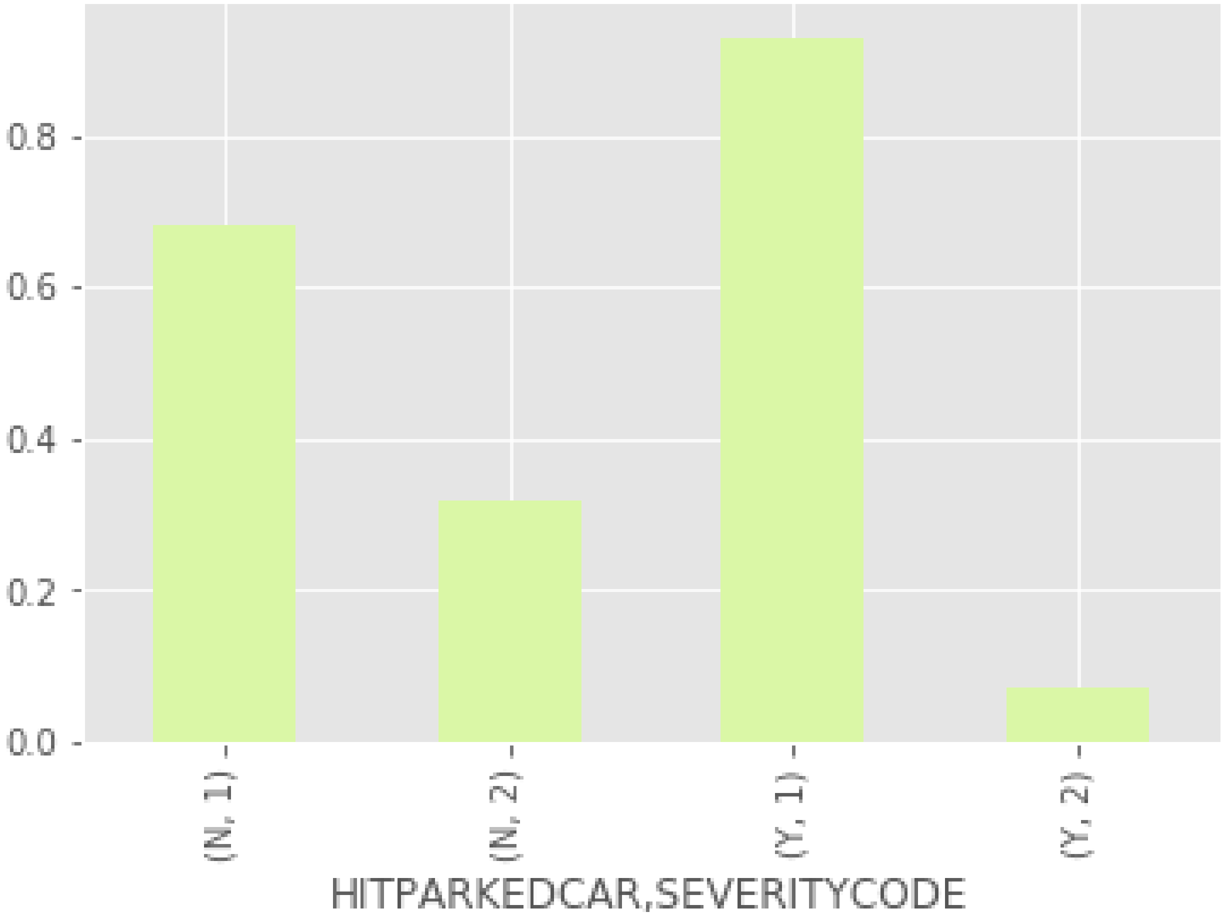


Figure 13 Hit Parked Car and Severity Code Relationship Analysis

3.2. Predictive Modelling

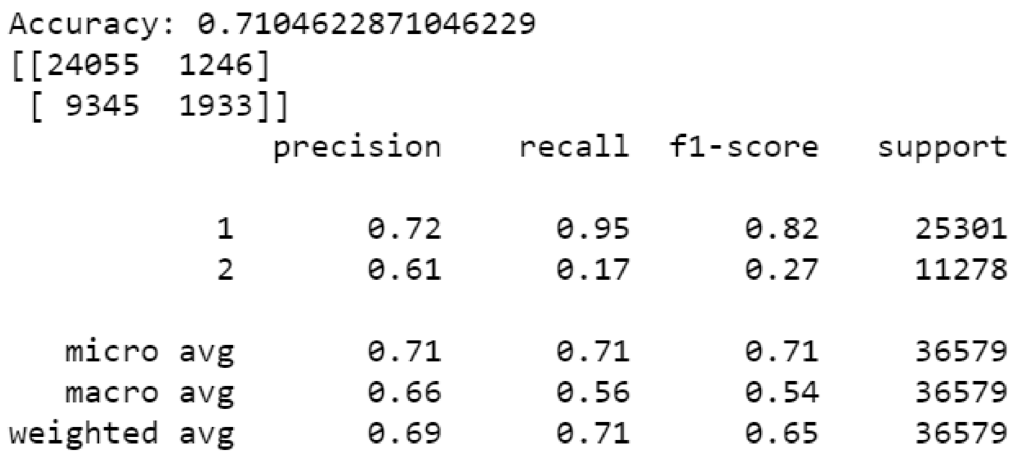
The data showed that when the incident involved hitting a parked car, majority of the time, the incident tend to have a severity code of 1. Also, we can see, as per data provided, data is covering all severity codes (only 1 and 2).

This problem is a classification problem with binary values as the target variables. The target variable; SEVERITYCODE is extracted as y while the remaining features are stored in the X data frame. The following supervised machine learning algorithms were applied to this problem to determine the best performing model that will be used to predict the severity code of an incident. For this model generation using following algorithms.

- Logistic Regression
- K-nearest neighbors
- Random Forest Classifiers
- Gaussian Naive Bayes Classifier
- Gradient Boosting Classifier

4. Results

- Logistic Regression

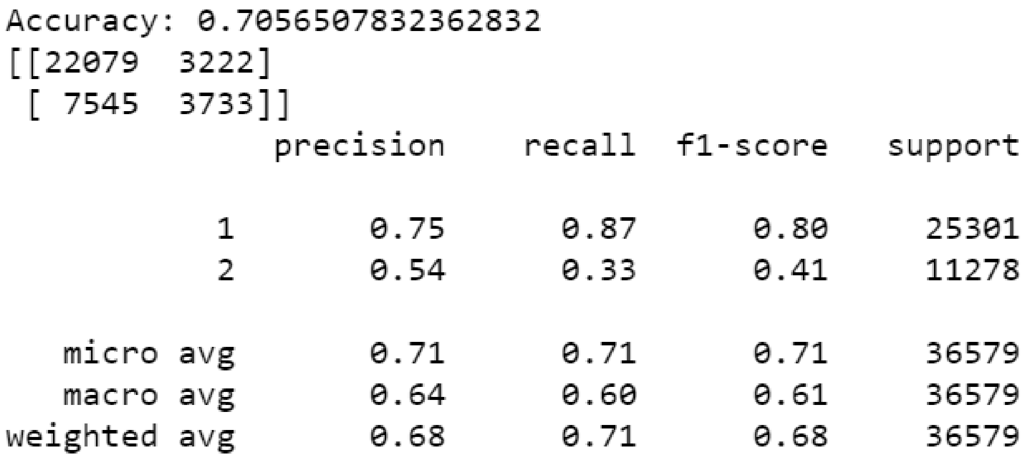


```
Accuracy: 0.7104622871046229
[[24055  1246]
 [ 9345  1933]]
              precision    recall  f1-score   support

           1       0.72      0.95      0.82     25301
           2       0.61      0.17      0.27     11278

   micro avg       0.71      0.71      0.71     36579
   macro avg       0.66      0.56      0.54     36579
weighted avg       0.69      0.71      0.65     36579
```

- K-nearest neighbors

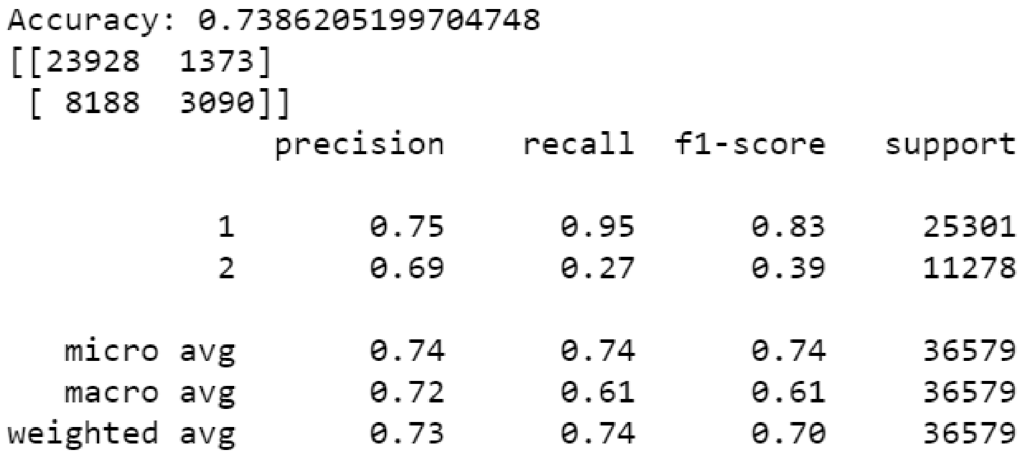


```
Accuracy: 0.7056507832362832
[[22079  3222]
 [ 7545  3733]]
              precision    recall  f1-score   support

           1       0.75      0.87      0.80     25301
           2       0.54      0.33      0.41     11278

   micro avg       0.71      0.71      0.71     36579
   macro avg       0.64      0.60      0.61     36579
weighted avg       0.68      0.71      0.68     36579
```

- Random Forest Classifiers

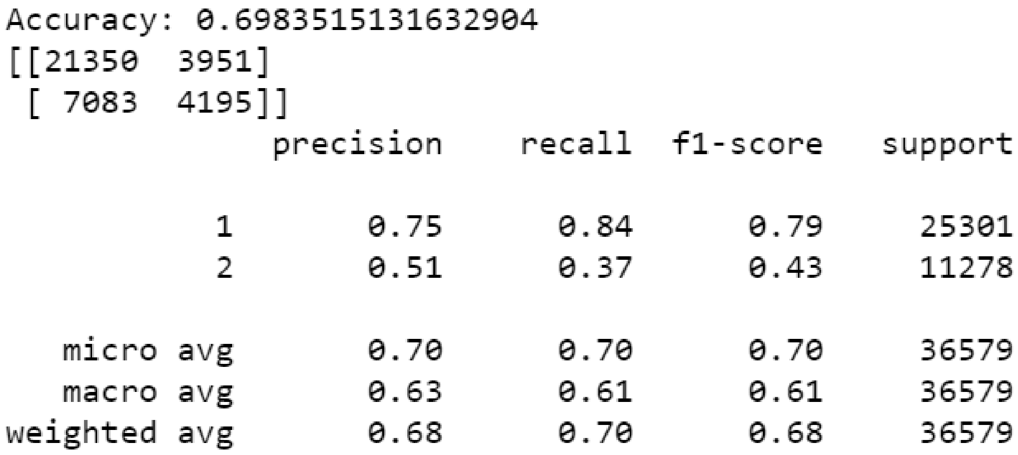


```
Accuracy: 0.7386205199704748
[[23928  1373]
 [ 8188  3090]]
              precision    recall  f1-score   support

           1       0.75      0.95      0.83     25301
           2       0.69      0.27      0.39     11278

   micro avg       0.74      0.74      0.74     36579
   macro avg       0.72      0.61      0.61     36579
weighted avg       0.73      0.74      0.70     36579
```

- Gaussian Naive Bayes Classifier

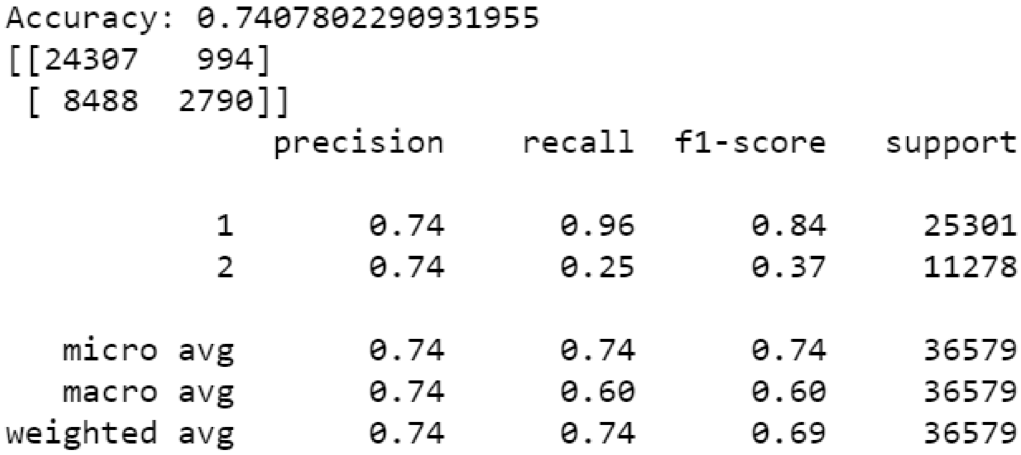


```
Accuracy: 0.6983515131632904
[[21350  3951]
 [ 7083  4195]]
              precision    recall  f1-score   support

           1       0.75      0.84      0.79     25301
           2       0.51      0.37      0.43     11278

   micro avg       0.70      0.70      0.70     36579
   macro avg       0.63      0.61      0.61     36579
weighted avg       0.68      0.70      0.68     36579
```

- Gradient Boosting Classifier

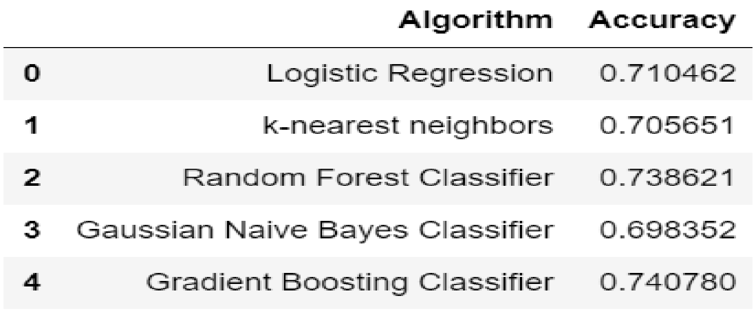


```
Accuracy: 0.7407802290931955
[[24307   994]
 [ 8488  2790]]
              precision    recall  f1-score   support

           1       0.74      0.96      0.84     25301
           2       0.74      0.25      0.37     11278

   micro avg       0.74      0.74      0.74     36579
   macro avg       0.74      0.60      0.60     36579
weighted avg       0.74      0.74      0.69     36579
```

The table below summarizes the performance of each of the model based on the accuracy of predicted result.



	Algorithm	Accuracy
0	Logistic Regression	0.710462
1	k-nearest neighbors	0.705651
2	Random Forest Classifier	0.738621
3	Gaussian Naive Bayes Classifier	0.698352
4	Gradient Boosting Classifier	0.740780

5. Discussion

From the table above it is observed that Gradient Boosting Classifier has the highest accuracy among used all algorithms and then next higher accuracy can be seen with Random Forest Classifier. This implies that the Gradient Boosting Classifier is the best model to use to predict the severity of the accident using the data provided.

6. Conclusion

In this project, I outlined the attributes that tend to affect the severity code od an incident such as weather, road condition, address type just to mention a few. I used different classification models to predict the severity code of an incident based on the attributes provided. The Gradient Boosting Classifier model proved to be the best model in making this prediction. This prediction will be helpful for residents as well traffic attendants and paramedics to predict the severity of incidents and plan in terms of providing medical attention and safety guidelines.