Predicting Car Accident Severity

Introduction

There are 280 million cars in the United States [1]. Such high level of traffic causes many accidents. In 2018, there were 12 million car accidents causing 36,560 deaths [1]. Several factors cause car accidents, such distracted driving, speeding, weather, drunk driving, etc.

To address this major problem, a machine learning model maybe developed to predict the severity of car accidents. Predicted car accident severity could help to warn drivers of potential risks and improve the preparedness of first responders.

Data

The data used in this project includes all types of collisions in Seattle from 2004 until 2020, with a total of 194,673 accidents. The data includes several attributes, such as accident location, address type (ally, block, or intersection), weather, road condition, speed, lane segment, etc. Attributes that are related to the causes of accidents, such as speeding and road conditions will be examined as they may help to predict the severity of an accident.

The dependent variable which is the severity of an accident has four levels:

- 0 refers to unknown.
- 1 refers to accidents with property damage.
- 2 refers to accidents with injuries.
- 2b refers to accidents with serious injuries.
- 3 refers to accidents with fatalities.

For this project, accident with severity level 0 (unknown) will be deleted as they do not contribute in training or testing our model to predict the severity of car accidents. To deal with imbalanced data, data will be under-sampled to balance the different level of accident severity.

Methodology

The data contains 194,673 accidents, with 136,485 (70%) belong to code 1 and 58,188 (30%) belong to code 2. Since there is good number of samples for the minority class, we will do class under-sampling to resolve the issue of class imbalance. Class under-sampling has the limitation of losing useful information.

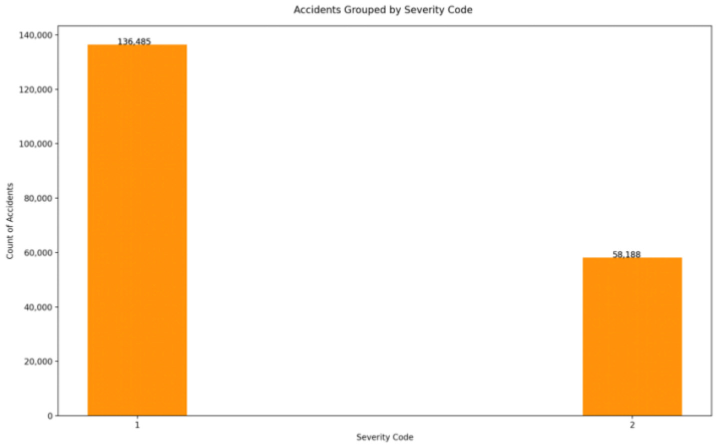


Fig1: Accidents grouped by severity code

The data contains about 37 features. Descriptive features, such as collision description, are not helpful to predicting their severity. Furthermore, to prevent data leakage, certain features that are related to the codes of accident severity, such as number of fatalities, must not be considered as they will defeat the purpose of prediction. Those features in addition to some duplicate features must be deleted. Certain features, such as speeding and whether an accident is intentioned, maybe useful to predict the severity of an accident. However, a lot of their values are missing, rendering them not useful.

We examine features to select those which have correlation with severity of accidents. The project will examine the features, show in Table 1, to predict the severity of accidents.

Seq	**Feature**	**Description**
1	Weather	Weather condition
2	Accident Day	Day of accident, such as Sun, Mon, Tue, etc.
3	Accident Month	Accident month, such as Jan, Feb, March, etc.
4	Accident Hour	Accident hour based on 24-hour format
5	Road Condition	Road condition
6	Light Condition	Light condition
7	Address Type	Address type which could be: Ally, Block, or Intersection
8	Collision Type	Collision type, such as head on, left turn, right turn, etc.
9	DUI	Driver is under the influence or not
10	Vehicle Count	Number of vehicles involved in the accident
11	Persons Count	Number of people involved in the accident
12	Pedestrian Count	Number of pedestrians involved in the accident
13	Bicycle Count	Number of bicycles if any

1. Weather

It may be expected that weather has a strong correlation with accident severity. However, based on the data, weather has no strong correlation, see Figure 2. When performing a Ch-squared analysis, Cramer's value is about 0.19 which shows weak relationship.

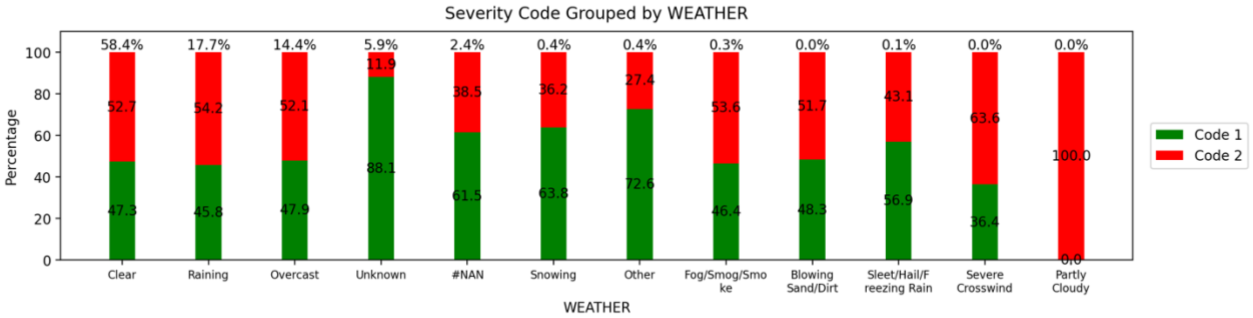


Fig2: Accidents grouped by weather condition

2. Accident Day

Accident day has no strong correlation with its severity, see Figure 3. When performing a Ch-squared analysis, Cramer's value is about 0.02 which shows weak relationship.

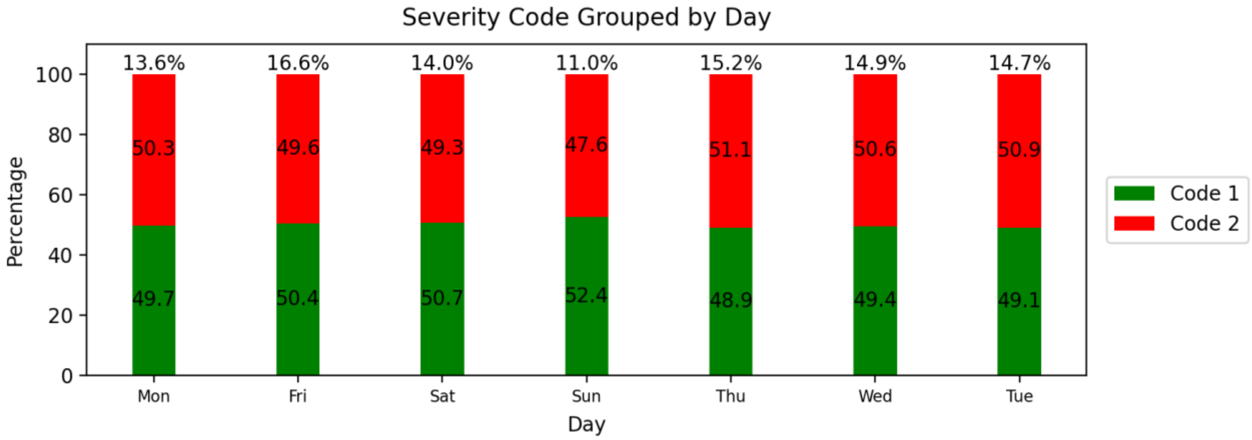


Fig3: Accidents grouped by day of the week

3. Accident Month

The same applies for accident month, it has no correlation with its severity, see Figure 4. When performing a Ch-squared analysis, Cramer's value is about 0.03 which shows weak relationship.

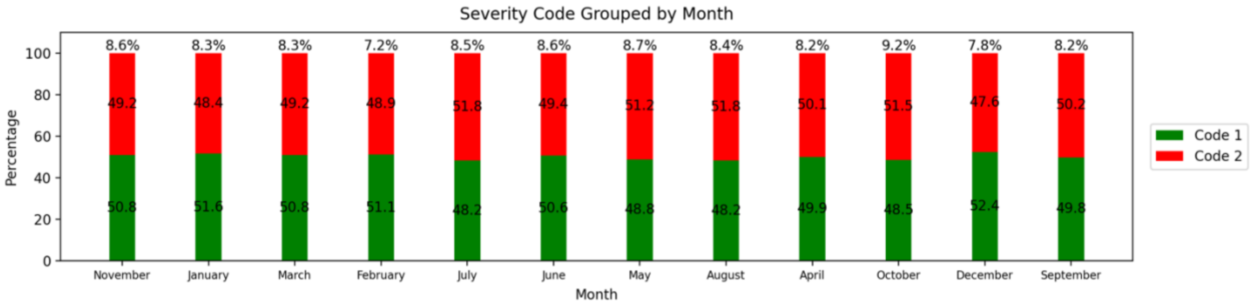


Fig4: Accidents grouped by month

4. Accident Hour

Accident time was missing from about 15.4% of accidents. Accidents with missing time were assigned the value of -1 as their hour. As shown in Figure 5, there seem no relation between accidents and their severity codes.

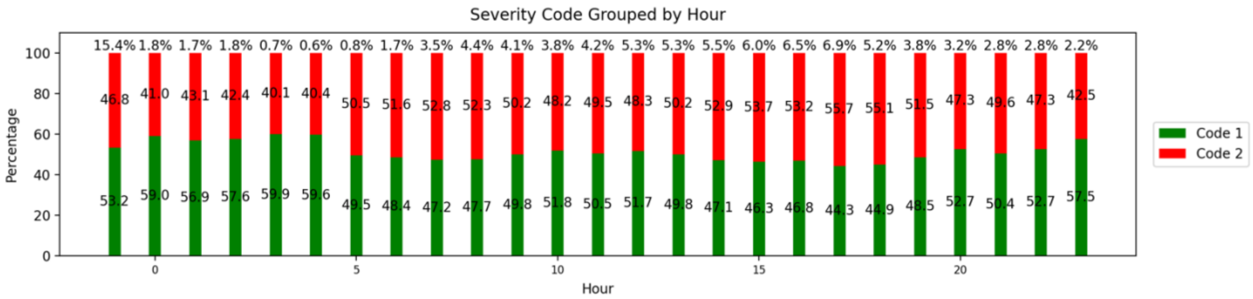


Fig5: Accidents grouped by time in hours

5. Road Condition

Road condition feature has some missing values for about 2.4% of accidents. We replace those values with the most frequent road condition which is Dry. There seem to be some correlation with the road condition and severity of accidents, see Figure 6. When performing a Ch-squared analysis, Cramer's value is about 0.21 which shows some relationship.

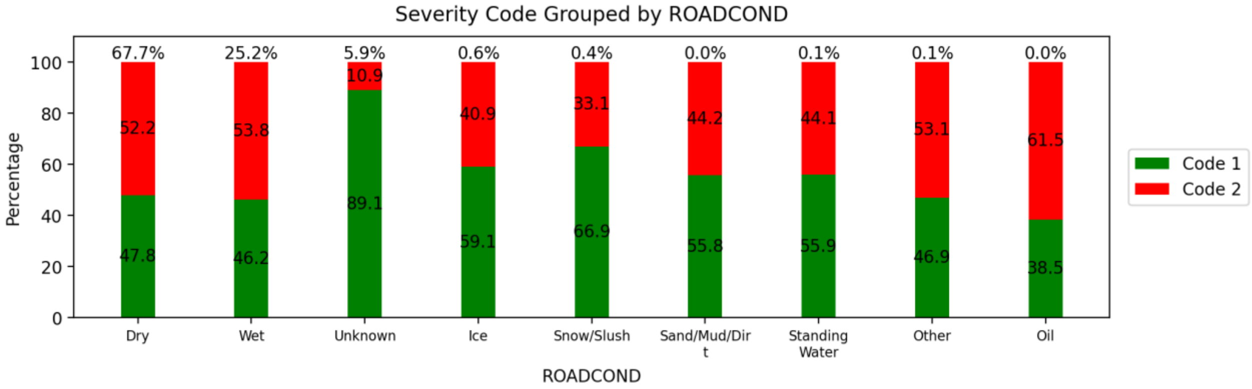


Fig6: Accidents grouped by road condition

6. Light Condition

Light condition feature has some missing values for about 2.4% of accidents. We replace those values with the most frequent light condition which is Daylight. There seem to be some correlation with the road condition and severity of accidents, see Figure 7. When performing a Ch-squared analysis, Cramer's value is about 0.2 which shows some relationship.

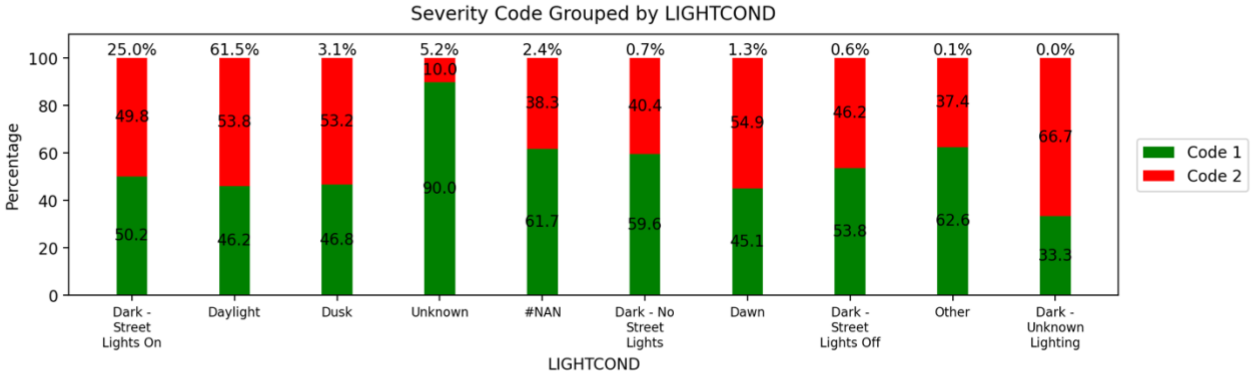


Fig7: Accidents grouped by light condition

7. Address Type

Address type feature has some missing values for about 0.8% of accidents. We replace those values with the most frequent address type which is Block. There seem to be some correlation with the address type and severity of accidents, see Figure 8. When performing a Ch-squared analysis, Cramer’s value is about 0.22 which shows some relationship.

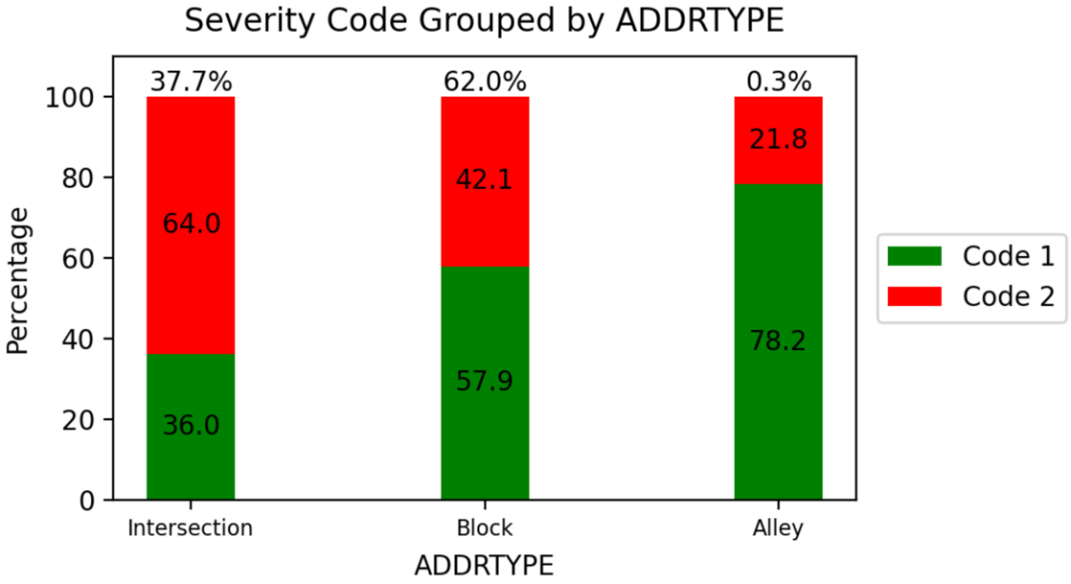


Fig8: Accidents grouped by address type

8. Collision Type

Collision type feature has some missing values for about 2.3% of accidents, which are deleted. There seem to be some correlation with the collision type and severity of accidents, see Figure 9. When performing a Ch-squared analysis, Cramer’s value is about 0.49 which shows very good relationship.

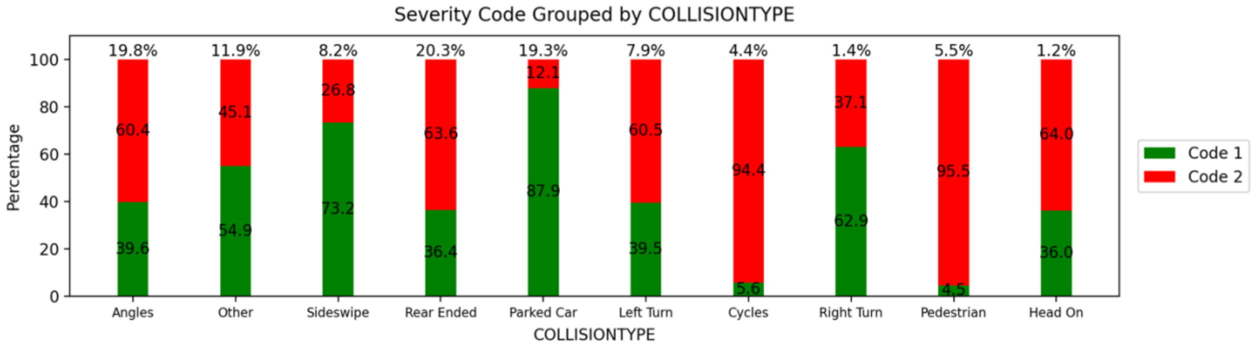


Fig9: Accidents grouped by collision type

9. DUI

DUI feature has a set of values, including [Y,N,0,1,NaN]. 0 and 1 are converted to N and Y. NaN values are replaced with the majority class which is N. There seem to be no correlation with the DUI feature and severity of accidents, see Figure 10. When performing a Ch-squared analysis, Cramer's value is about 0.05 which shows weak relationship.

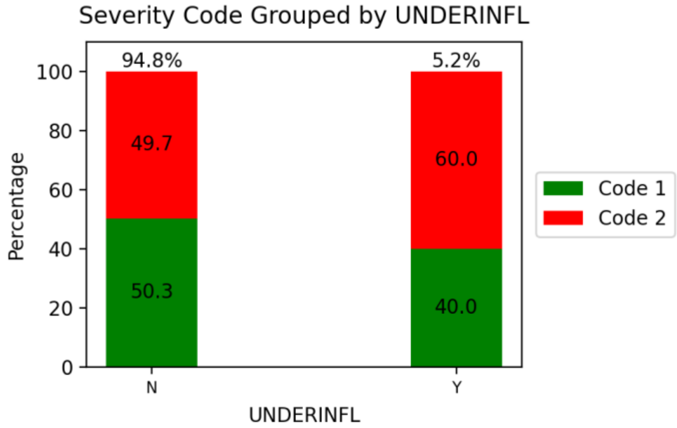
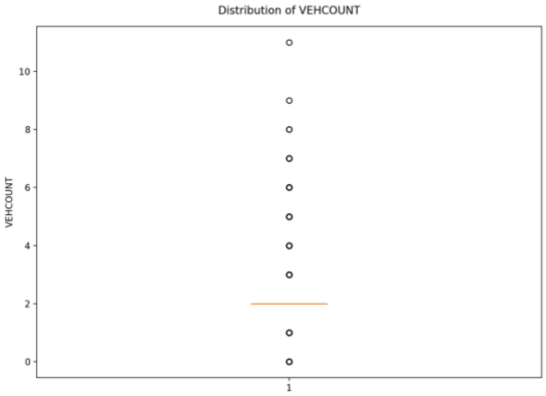


Fig10: Accidents grouped by DUI

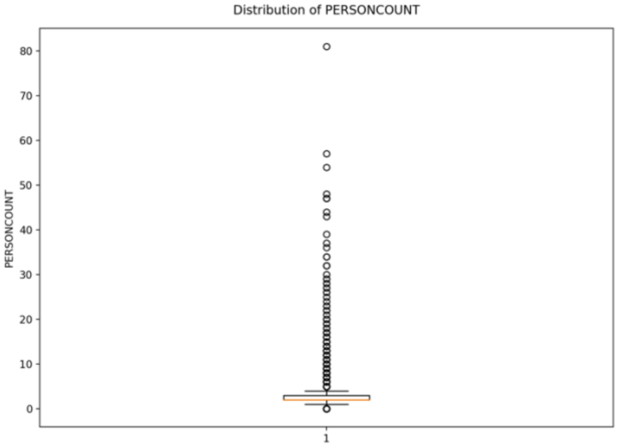
10. Vehicle Count

It is clear from the boxplot below that accidents 0 and with 5 or more vehicles are outlier. As a result, those accidents will be deleted.



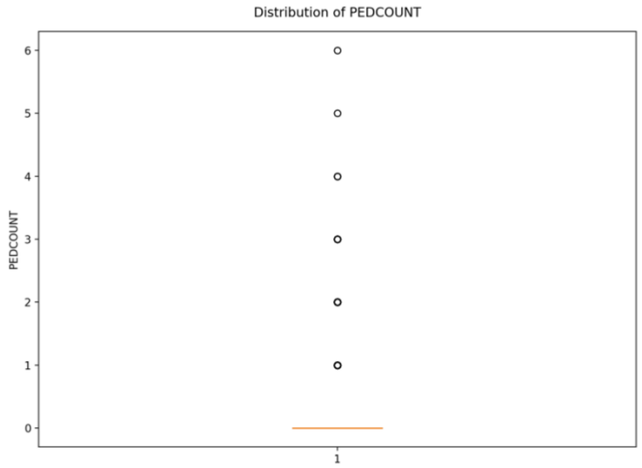
11. Persons Count

It is clear from the boxplot below that accidents with about 9 or more people are outlier. As a result, those accidents will be deleted.



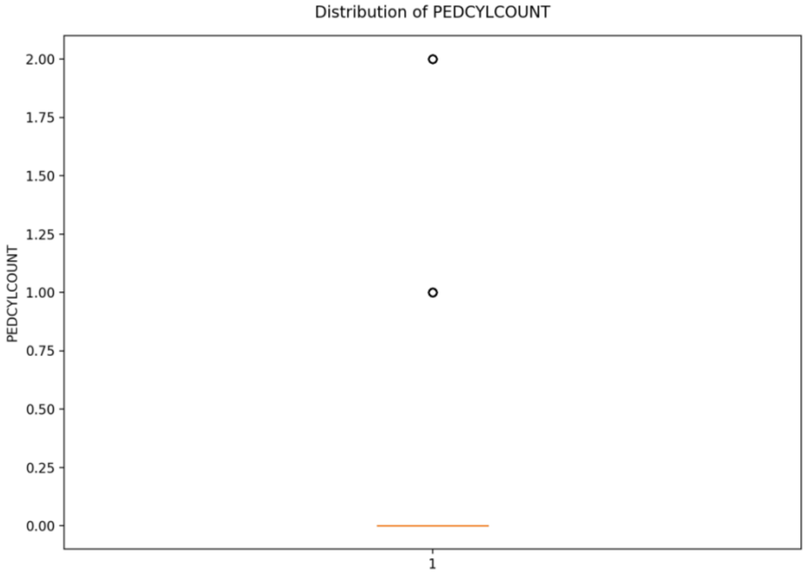
12. Pedestrian Count

It is clear from the boxplot below that accidents with about 3 or more pedestrians are outlier. As a result, those accidents will be deleted.



13. Bicycle Count

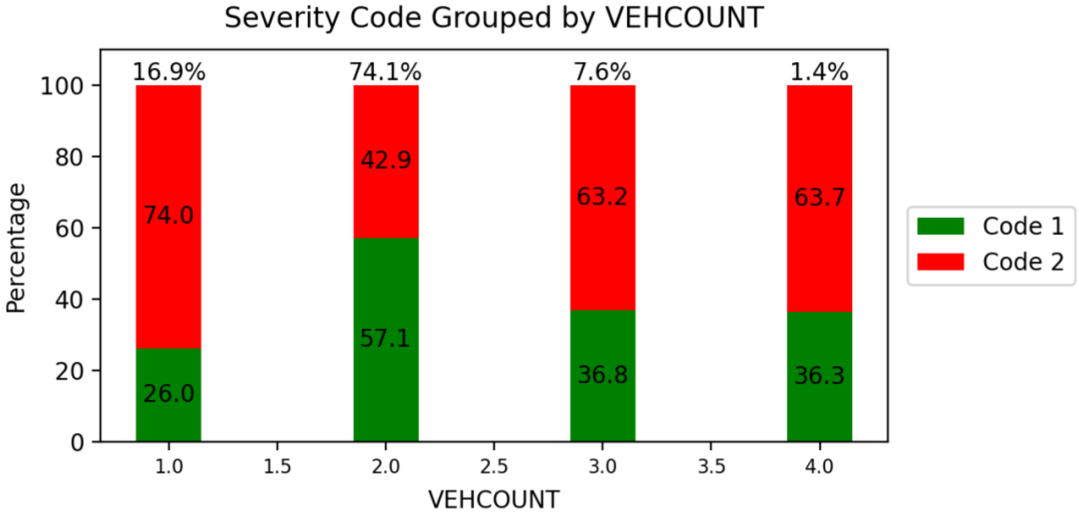
It is clear from the boxplot below that accidents with more than 1 bicycle are outlier. As a result, those accidents will be deleted.



The table below shows the correlation between numeric variables.

```
                 INCKEY  PERSONCOUNT  PEDCOUNT  PEDCYLCOUNT  VEHCOUNT  \
INCKEY         1.000000    -0.066338  0.034381     0.043216 -0.027777
PERSONCOUNT   -0.066338     1.000000 -0.053523    -0.060960  0.430674
PEDCOUNT       0.034381    -0.053523  1.000000    -0.049164 -0.394505
PEDCYLCOUNT    0.043216    -0.060960 -0.049164     1.000000 -0.357746
VEHCOUNT      -0.027777     0.430674 -0.394505    -0.357746  1.000000
SEVERITYCODE   0.030872     0.150489  0.220200     0.187279 -0.096880
```

Looking at the severity code row, we see about 0.15, 0.22, 0.19 correlation with Person Count, Pedestrians Count, and Bicycle Count respectively. There is a – 0.97 negative correlation between number of vehicles and severity of accidents, which is counter-initiative. This relationship is depicted in the below figure.



- **Feature Selection**
  The exploratory analysis performed in the previous section gave a good idea about the features of the data. We pick those features which have correlations with the severity of accidents. The selected features are shown in the below table.

Seq	Feature	Description
1	Road Condition	Road condition
2	Light Condition	Light condition
3	Address Type	Address type which could be: Ally, Block, or Intersection
4	Collision Type	Collision type, such as head on, left turn, right turn, etc.
5	Vehicle Count	Number of vehicles involved in the accident
6	Persons Count	Number of people involved in the accident
7	Pedestrian Count	Number of pedestrians involved in the accident
8	Bicycle Count	Number of bicycles if any

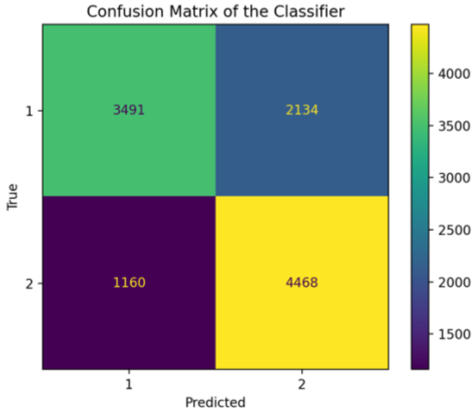
Evaluation

We use Random Forest algorithm with the selected features to predict the severity of accidents. Random Forests is a supervised learning technique that employs an ensemble of decision trees that can be used to solve classification and regression problems. It has the advantage of avoiding overfitting.

First, will perform one hot-encoding on categorical features. Data is randomly split into 90% training and 10% testing. For training, we perform 10-cross fold validation and grid search to set the model parameters, such as number of trees and max depth. For evaluation, we show the confusion matrix and report the accuracy and F1-Score.

Results

For training, we focused on the F1-Score which achieved 68%. Testing results are 67% and 70% for F1-Score and accuracy, respectively.



Discussion

The model achieved 68% in F1 score and 70% in accuracy. The model is not able to achieve good results because of the weak correlation between features in the data and severity of accidents. Certain features, such as speeding, may have strong correlation with accident severity, but it is missing for majority of accidents. Trying additional models may not provide significant improvements.

Conclusion

This report addresses the problem of predicting car accident severity to warn drivers and first responders. The data is obtained from Seattle Department of Transportation. The report explores existing features and evaluates their correlation with car accident severity. A Random Forest model is implemented to predict the severity of accidents, achieving an F1 Score of 67% in testing data.

References

1. https://www.statista.com/topics/3708/road-accidents-in-the-us/