Predict Severity of a Traffic Accident

1. Introduction

1.1. **Background**

There are many uncertainties on the road that could happened at any time to anyone. Those uncertainties could lead to a traffic accident. For example, bad weather, speed, location, and human error determine how the accident is. But is there any correlation between all those uncertainties factors with the severity of an accident? Imagine, if we could predict the severity of an accident by the location, weather, road, time, etc., then the driver could be more aware or more careful or take another road or cancel the plan.

1.2. **Problem**

The projects aim to predict the severity of an accident based on the road condition, weather, time, etc.

1.3. **Interest**

Government, traffic police, and driver could use the data to predict the severity of traffic accident and decide what to do faster.

2. Data

2.1. Data Sources

The data that is used in this project is collision data from Seattle Government from 2004 till present. The data contains 38 features, 194673 rows, with a categorical label that consists of 1 and 2 where 1 represents property damage and 2 represents injury. For more information about the data, visit here.

2.2. Data Cleaning

There are several problems with the dataset and the most important one is the NaN values because we cannot analysis the data with NaN values in it. In this section, the data in certain columns will be deleted, filled, or replace based on its value.

First the filled data. The columns that consist of NaN data in categories will be filled with ‘unknown’ because I will transform the categories text data to one hot encoding. Second is the deleted data. For latitude and longitude data, there are about 5,000 NaN values. Since I can’t replace the data with anything, I decided to delete the row. Third is the replaced data. Some features contain some strange values. Like ‘Underinfl’ consists of ‘Y’, ‘N’, nan, ‘0’, and ‘1’, so I changed ‘0’ and ‘1’ consecutively to ‘N’ and ‘Y’.

2.3. Feature Selections

In this part, I delete some features and most of them because have a lot of NaN values, a description that has already been represent in other features, and irrelevant based on the correlation value.

3. Exploratory Data Analysis

3.1. Label Data

Label data consists of 5 categories which are

Table 1. Label data description

Categories	Description	Count
1	Property Damage	137776
2	Injury	58842
2b	Serious Injury	3111
3	Fatality	352

and this bar graph below show the count of each categories.

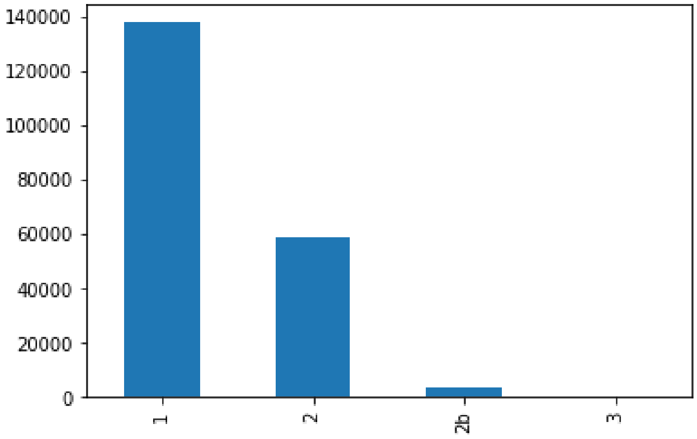

Figure 1. The count value of each category in label data.

Based on the table and figure above, we know that the label data is imbalanced, I will deal with it using classification report to represent the accuracy.

3.2. Relationship Between Features and Severity Code

In this boxplot is shown the relationship between features and severity code.

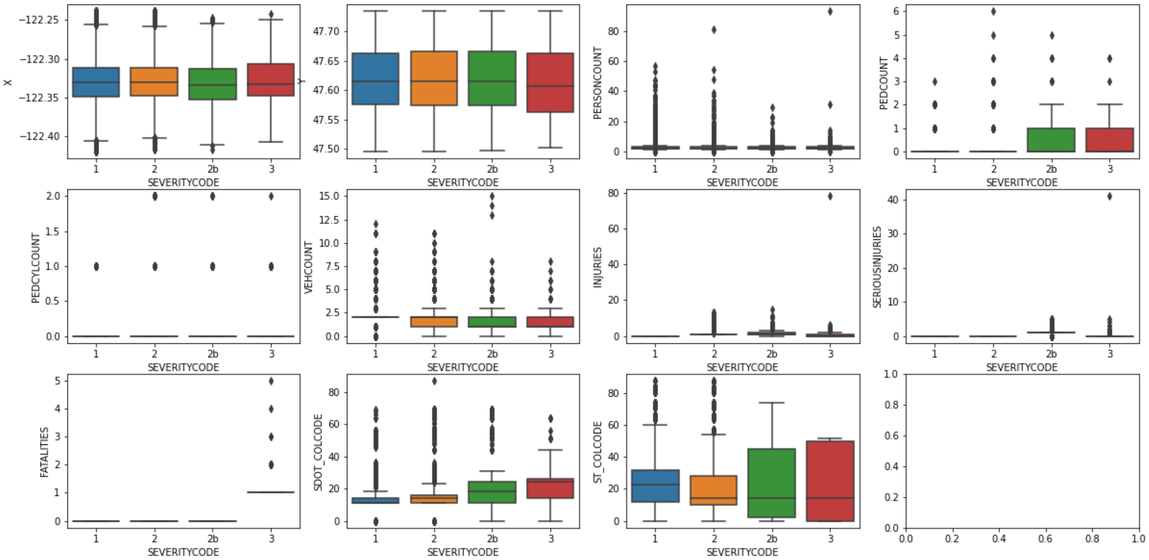

Figure 2. Boxplot that shows the relationship between categorical features

4. Method

In this project I am using One Hot Encode, Random Forest, and Neural Networks.

4.1. One Hot Encode

This method is to transform a categorical data to several columns of data with 0 and 1 values. Since the data has a lot of categorical features, this method is useful. After the data is transformed, the features of data became 58 features in total.

4.2. Split Train and Test

Split train and test data are a method to avoid overfitting. I am using a scikit learn package with test size equals to 0.65 and shuffle rows is true.

4.3. Random Forest

Random forest is a classification method based on tree model. Random forest is basically a decision tree with batch of data to make the method run faster.

In this section, I am using a random forest package from scikit learn. I also use randomized search cv to tune the hyperparameter.

The result of the prediction is shown in the table below.

Table 2. Classification report

Categories	Precision	Recall	F1 Score
1	1	1	1
2	1	1	1
2b	0.99	0.99	0.99
3	1	0.85	0.92
Accuracy			1
Macro avg	1	0.96	0.98
Weighted avg	1	1	1

4.4. Neural Networks

Neural networks is one of machine learning method that could be use to predict data. The method is inspired by how our brain works.

In this section, I am using a package from Keras with one hidden layer consists of 256 neural. I am using MSE as the loss function, Adam as the optimizer, and Relu as the activation function.

The result of the prediction is shown in the table below.

Table 3. Classification report

Categories	Precision	Recall	F1 Score
1	1	1	1
2	1	1	1
2b	1	0.99	0.99
3	1	0.99	1
Accuracy			1
Macro avg	1	0.99	1
Weighted avg	1	1	1

5. Conclusion

In this project I analyzed a data of severity of traffic accidents at Seattle from 2004 till present using Random Forest and Neural Networks. Based on the classification report table, shown that neural networks gave a better result than random forest classifier especially for category '3'. But in conclusions, both methods are given a good result to predict future traffic accidents.

6. Future Directions

In this project, I am only using two classification models, however there are still many models that I haven't tried that might fit best for this problem.