Table of Contents

Contents

1.0 Introduction 4

2.0 Data 4

2.1 Data Cleaning 4

2.2 Feature selection 4

3.0 Methodology 5

3.1 Exploratory Analysis 5

3.2 Machine learning method selection 5

4. Results and Discussion 5

4.0 Conclusions 5

List of Figures

No table of figures entries found.

List of Tables

Table 1. Selected features and code 4

1.0 Introduction

In an effort to reduce the frequency of car collisions in a community, an algorithm must be developed to predict the severity of an accident given the current weather, road and visibility conditions.

2.0 Data

This section was used to describe the data that will be used to solve the problem and illustrate the source of the data. There are totally 38 attributes and 194673 data rows in this dataset, and it includes all types of collisions from 2004 to present.

2.1 Data Cleaning

To predict accident severity, the code of the target variable 'SEVERITYCODE' that used to measure the severity of an accident, has been changed to the value of 1, which means property collision but no injury, and 2, indicating injury collision.

2.2 Feature selection

The feature attributes used to weigh the severity of an accident are 'WEATHER', 'ROADCOND' and 'LIGHTCOND'. In order to analysis and predict the severity of accident, the features should be encoded to numerical type.

Table 1. Selected features and code

Selected Feature	Code
'WEATHER'	Weather Conditions (0 = Clear, 1 = Overcast and Cloudy, 2 = Windy, 3 = Rain and Snow)
'ROADCOND'	Road Conditions (0 = Dry, 1 = Mushy, 2 = Wet)
'LIGHTCOND'	Light Conditions (0 = Light, 1 = Medium, 2 = Dark)

3.0 Methodology

3.1 Exploratory Analysis

The feature of 'SEVERITYCODE' has been selected as a target variable, the attributes used to weigh the severity of an accident are ‘WEATHER', 'ROADCOND' and 'LIGHTCOND'. After dropping of the missing data, there are totally 189337 data used for this model. In the beginning of this notebook, we had categorical data that was of type 'object'. This is not a data type that we could have fed through an algoritim, so label encoding was used to created new classes that were of type int8; a numerical data type. After solving that issue we were presented with another - imbalanced data. As mentioned earlier, class 1 was nearly three times larger than class 2. The solution to this was downsampling the majority class with sklearn's resample tool. We downsampled to match the minority class exactly with 58188 values each.

3.2 Machine learning method selection

Following models have been selected to fit this dataset, K-Nearest Neighbor (KNN), Decision Tree and Logistic Regression. KNN will help us predict the severity code of an outcome by finding the most similar to data point within k distance. A decision tree model gives us a layout of all possible outcomes so we can fully analyze the concequences of a decision. It context, the decision tree observes all possible outcomes of different weather conditions. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable

4. Results and Discussion

Three ML models, K-Nearest Neighbor, Decision Tree and Logistic Regression, were selected for this dataset. Although the first two are ideal for this project, logistic regression made most sense because of its binary nature. Evaluation metrics used to test the accuracy of our models were jaccard index, f-1 score and logloss for logistic regression. Choosing different k, max depth and hyparameter C values helped to improve our accuracy to be the best possible.

4.0 Conclusions

Based on historical data from weather conditions pointing to certain classes, we can conclude that particular weather conditions have a somewhat impact on whether or not travel could result in property damage (class 1) or injury (class 2).