Capstone Project - Car accident severity

1. Introduction

More than 1.2 million people worldwide die each year in road traffic accidents (RTAs) and another 20-50 million are injured. The rise in the number of fatalities in road traffic accidents is an increasingly serious problem. Indeed, the number of road traffic participants and the number of vehicles in the world are growing every year. Of course, the main factors in road accidents are driving under impact of alcohol and drugs and speeding. These causes of accidents are prevented by means of education for road users and fines. In turn, the negative impact of bad weather conditions, lighting and road quality can be analyzed and preventive measures (information board) for participants can be taken to prevent accidents.

2. Business Background

The main purpose of road accidents analyzing is to save human lives. Hence, the task of collecting, processing and analyzing data on road traffic accidents is important for the following reasons:

- reduction of fatal accidents to save human lives
- reduction in the total number of accidents reduces the number of payments for a number of insurance policies
- reduces the cost of hours of work for police and other rescue services
- reduces the number of traffic difficulties caused by an accident.

In this project only an impact of wether, road and light conditions on collision are investigated.

3. Data

Accident registration data is provided by the City of Seattle and has been recorded since 2004. Data is updated weekly (https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv). The data is presented as a table in hard currency format and contains 38 columns and 194673 lines. The target parameter is an accident "severity" in terms of human fatality, traffic delay, property damage, or any other type of accident bad impact. To predict the class of accident, the following parameters of the road traffic situation were selected:

- *SEVERITYCODE* (a code that corresponds to the severity of the collision: 2 —injury, 1—prop damage);
- *WEATHER* (A description of the weather conditions during the time of the collision)
- *ROADCOND* (The condition of the road during the collision)
- *LIGHTCOND* (The light conditions during the collision)

4. Data Preprocessing

All data are checked for the presence of missing information. Any rows with missing information will be removed from the data frame to create a Machine learning model.

1. Remove all severity data if driver involved was under the influence of drugs or alcohol or speeding or inattention were factors in the collision. There are 4 values for driving under the influence of alcohol and drugs. In this case it was decided to consider 0 and H as 0 (there is no influence of alcohol or drug intoxication) and 1 and Yes as 1.

2. Remove all rows with «Unknown» and «Other» categories from data frame.

3. Remove all NaN rows.

4. Replace text in string format which describes wether, road and light conditions with a series of binary elements.

5. Balance data frame.

```
[35]: dftd['SEVERITYCODE'].value_counts()

Out[35]: 2    40857
         1    40857
         Name: SEVERITYCODE, dtype: int64
```

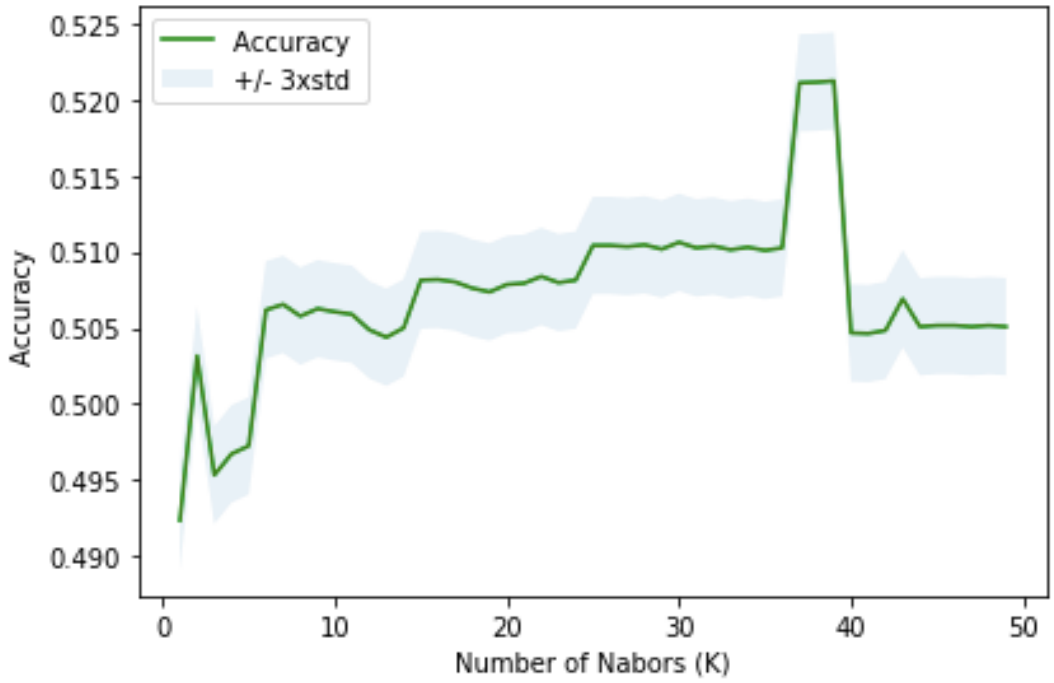
After Balancing

6. Normalize data frame.

5. Methodology

After the input data has been analysed and processed, you can proceed to the choice of a machine learning model. To train the model, the data is broken down into training data and test data. The data is split 70% for training and 30% for testing.

1. The K-Nearest Neighbours (KNN) algorithm uses feature similarity to predict the values of new data points, which also means that a new data point will be assigned a value based on how closely it matches points in the training set. The model is tested for K from 1 to 50 and the optimal value is K = 39.



Accuracy of KNN for different *k*

2. Decision tree training is a technique whose purpose is to create a model that predicts the value of a target variable based on some input variables. Decision trees can also be described as a combination of mathematical and computational techniques to describe, categorise, and generalise a given set of data. The tree can be "trained" by splitting the set into subsets based on checking the attribute values. This process, which repeats recursively on each subset obtained, is called recursive partitioning. Recursion stops when a subset at a node has the same target variable value, or when splitting adds no value to the predictions.

3. The main idea of logistic regression is that the space of initial values can be divided by a linear boundary (ie, a straight line) into two areas corresponding to the classes. This boundary is set depending on the available input data and the training algorithm. For this to work, the original data points must be separated by a linear border into the two aforementioned areas.

4. Support vector machine. The main idea of the method is to construct a hyperplane that separates the sampled objects in an optimal way. The algorithm works under the assumption that the greater the distance (gap) between the dividing hyperplane and the objects of the shared classes, the smaller the average error of the classifier will be.

All 4 models are suitable for the classification tasks that we have in this work: an accident without risk to human health or not.

6. Results

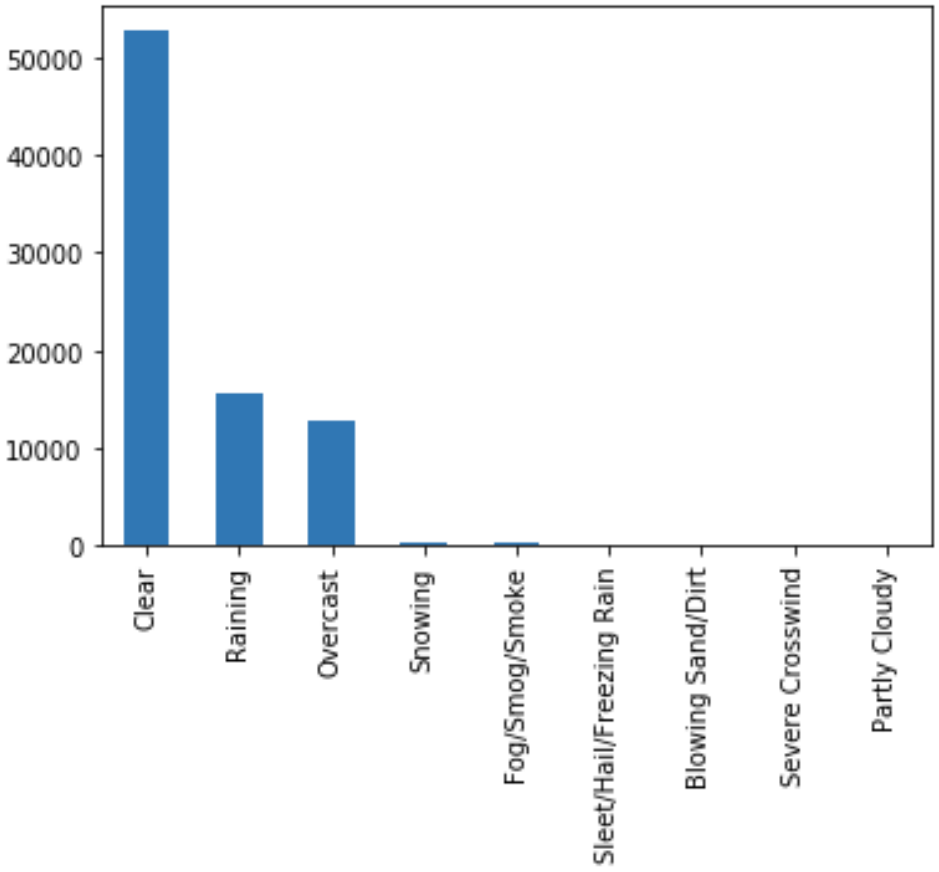
The table below lists the results for four machine learning models. It can be seen from this table that the highest accuracy for this task is demonstrated by the logistic regression models and the decision tree method.

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.52	0.48	NA
Decision Tree	0.52	0.50	NA
SVM	0.52	0.49	NA
LogisticRegression	0.52	0.50	0.69

Accuracy results of four ML models

7. Discussion

Machine learning methods can be widely used for tasks such as predicting accidents based on a number of parameters.In our case, this is a classification problem - «Yes» or «No» (1 or 0). Having made a forecast of the probability of accidents in accordance with the parameters of weather, lighting and road conditions, you can take preventive measures in relation to road users. Although, after analysing the ratio of the number of accidents to weather conditions, it can be seen that most accidents occur in good weather and at dry road. There are more days with good weather and dry



Number of accidents for different weather conditions

roads in general. It follows that the human factor prevails over the factor of external parameters such as weather, road quality and lighting. Unfortunately, data for accidents because of alcohol or drug intoxication, inattention of driver or speeding have significant gaps. Improving the collection of data on these parameters will help predict accidents.

Also, the accuracy of the results of machine learning models in this work can be improved by further changing the parameters of the models.

8. Conclusion

This project analyzed road accidents in the city of Seattle from 2004 to 2019. In particular, using machine learning methods, models have been built to predict the quality of accidents (with a threat to human health or not) depending on weather conditions, road conditions and the level of illumination. The highest accuracy for this task is demonstrated by logistic regression models and the decision tree method. The accuracy of the models can be improved by varying the model parameters. Based on these results, warning systems for drivers of high risk of accidents can be developed. Also, observations were made based on data analysis that most of the accidents occur in good weather conditions, good lighting and a dry road. It follows that other driving factors such as drunk driving, inattention while driving and speeding may have a greater role in the cause of accidents. To take into account the impact of these parameters, it is necessary to keep a more careful record of these parameters when registering an accident.

Thank you for reading!