Coursera – Applied Data Science Capstone

Roman Semin

“Car Accident Alert System”

Business Problem

Cars and motorcycles are the most dangerous way of transportation resulting in around 219.85 fatalities per billion passenger miles according to Statista [1]. In the US over 6 million people get into car accidents every year, and on average 90 people die every day as a direct or indirect result of them [2]. Most often the accidents occur because of conditions that prevent drivers from communicating together well on the road, including bad weather, road and light conditions. Considering that these are the factors not controllable by drivers who take a certain path, it makes it extremely dangerous to drive when such conditions are bad. Conditions like slippery roads, heavy fog, or even rain take away control from the driver, putting lives of people on the road of a roulette, which many people lose and end up either in a hospital or dead.

The purpose of the project is to find out the correlation between conditions such as road, weather or light conditions, and severity of potential accidents that might happen. The report is answering two questions.

The first question is how likely you are to get into an accident.

The second question is what type of accident you are most likely to get into depending on the conditions mentioned above.

Data Description

Through inspection it was found that the dataset contains 38 columns. Most of the columns are qualitative (ex. text, object). Each of 194673 records describes details about the accident recorded by police. The data that needed for analysis of the problem included four columns: 'ROADCOND', 'WEATHER', 'LIGHTCOND', 'SEVERITYCODE'. For the ‘ROADCOND’ column, the data is significantly skewed (60% of the dataset) towards *Dry* value. The dataset also contained ~15’000 unknown values, which had to be removed. For ‘WEATHER’ ‘LIGHTCOND’ columns, similar observation as 60% of data is skewed towards one value such as *Clear* and *Daylight* respectively. For all records, the 'SEVERITYCODE' has ⅔ ratio of a minor accident and only ⅓ values for accidents that caused some traumas. It is quite an amount. The approach to solving the problem is finding the relationship in severity code increase and one of the conditions. The data from the three road, weather and light condition columns would be inspected to find the best combination with the highest probability of the severity code and develop a machine learning model that based on the conditions predicts the severity code of the accident.

Methodology

This section is divided into three parts to make the report more structured. The first section is preparing the data for analysis. It is an important step to take only meaningful data from the provided dataset before the analysis and machine learning training. The second section is data analysis where the high level of relationship discovery is done and the trends are visualized. Last, and the third section of the report is applying correct machine learning algorithms and evaluating the training result of each algorithm.

Data Preparation

As mentioned previously, in the preparation stage, the focus is to prepare clean data to work with in later sections. First step was to identify the right columns to work with. The problem talks about effects of the conditions on the chance and severity of the accident on the road, thus four columns were picked for further analysis: 'ROADCOND', 'WEATHER', 'LIGHTCOND', 'SEVERITYCODE'. Columns 'ROADCOND', 'WEATHER', and 'LIGHTCOND' correspond to the uncontrollable conditions person might face while driving a vehicle. THe columns 'SEVERITYCODE' corresponds to the severity of the accident.

As the data used in these columns is mostly qualitative, the columns were inspected for the values contained in them. During the inspection, the dataset turned out to have at least 10% of records with at least one of the target condition columns were unidentified. These records were cleared for the sake of having perfectly clean data.

After conducting cleaning of empty records and limiting the target dataset only to four columns, the total shape of the data became (169949, 4) and it's more than 80% of all original records.

Analysis

The focus of this step in the project was to understand high level relationships between columns. Visualizing distribution of values [see Fig. 1] in the columns gave an understanding, what data to work with and what biases to consider. The data was mostly plotted in bar graphs, pie charts with percentage distribution and histograms as these are the most descriptive plotting methods for categorical data. Seeing the distribution, I had an understanding how to proceed with applying machine learning algorithms.

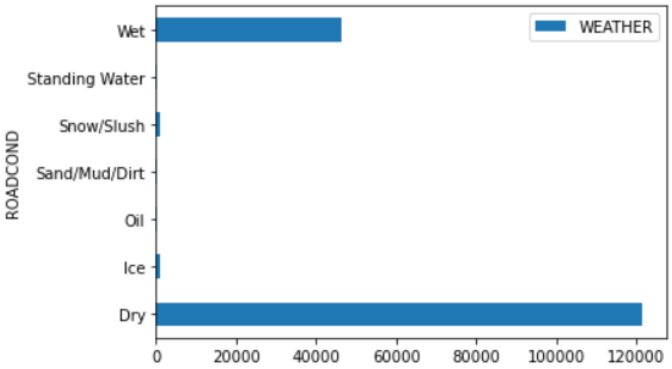

Figure 1. WEATHER column distribution

Machine Learning

With a good knowledge of data machine learning algorithms were applied. Three machine learning models were chosen to be tested: logistic regression, KNN and decision tree. As all data is categorical and the problem requires to predict the severity code (already pre-defined) of an accident, only classification algorithms could have been used for the problem. The three machine learning models mentioned previously are the best candidates for classification.

For each of the models, the split ratio into test and train data was configured to be 30% by default. The confusion matrix with four folds was also used to test every model.

As for measuring accuracy of the models, MSE and R1 scores were taken for performance of every model.

KNN

K-nearest-neighbour is a great algorithm when we want to classify a record based on a quantitative metric, thus the analysis was not as precise as intended. As a random test sample was chosen to be 30% of the dataset and the neighbours were measured based on Minkowski's distance. KNN did not use the grid search and the value k of nearest neighbors was chosen to be 4 as a default.

Decision Tree

The most applicable machine learning algorithm to predict the outcome. The decision tree is the most applicable algorithm for classifying categorical data and in case of conditions affecting the severity of the accident it was very applicable.

Logarithmic Regression

Logistic regression does not seem to be an algorithm to be applicable for the problem since it can only classify the record in two groups. Despite severity code having more five options, for the clean dataset there were only two options of severity code: 1 or 2, thus the logistic regression was applied.

Results

Evaluating the models with confusion matrix and grid search the best accuracy was show by the decision tree model (67% mean). The rationale for this result is as discussed in the previous section, that the decision tree is the best way to classify data based on solo categorical data from different columns. Other models such as KNN and logistic regression showed worse performance (63.2% and 56.1% respectively) partially because they were not properly adapted to the data. Logistic regression also prefers to classify the data based on qualitative data, but it performs worse, with no current explanation, as it predicts that all severity code values should be classified as 1, and since on average there are more 1 severity code data points, the ~50% accuracy is understandable but not acceptable.

Discussion

Thinking logically, road conditions can be connected to certain locations can be more severe in certain parts of Seattle where the construction was happening. For future analysis, we can find the relationship between road condition severity and location (longitude and amplitude), because if there is a direct relationship, the location can be used as a quantitative metric for the KNN model to increase its accuracy.

In the future modifications to the models, KNN should also use grid search as it would allow to find the most optimal value of k for accuracy of the model.

Lastly, the decision tree should be considered to be adapted to a random forest algorithm as it is proven to be faster than the convention decision tree and it also incorporates linear regression improving accuracy of the model.

Conclusion

In conclusion the model with the highest accuracy is a decision tree that predicts severity of the accident based on the road, weather and light conditions with 67% accuracy. The algorithm is not very reliable but it has enough accuracy to alert drivers based on the path they choose to take, if they are likely to get into the accident taking weather and light conditions throughout the whole path.
The project was a success and the next steps would be to develop an interface for the alert system and also to improve the model based on recommendations to rise accuracy over 80% minimum.

[1] M. Armstrong and F. Richter, “Infographic: The Most Dangerous Ways to Travel in the U.S.,” *Statista Infographics*, 04-Jun-2019. [Online]. Available: https://www.statista.com/chart/18264/the-most-dangerous-ways-to-travel-in-the-us/. [Accessed: 29-Sep-2020].

[2] “Car Accident Statistics in the U.S.: Driver Knowledge,” *DriverKnowledge*, 25-May-2019. [Online]. Available: https://www.driverknowledge.com/car-accident-statistics/. [Accessed: 29-Sep-2020].