Data Science - Professional Certificate

Coursera - IBM - Capstone Project - Car accident severity

Ricardo Yoshitaka Ikeda

August 31, 2020

Introduction

Business Understanding

The entity that compiled the data is a US state government agency. Seattle Department of Transportation.

We can draw some conclusions from the problems that occur when accidents involving cars and people occur:

- there may be people with injuries or fatalities;
- displacement of police vehicles;
- displacement of rescue vehicles;
- congestion in the area;
- delays for car and bus users.

When analyzing accident information, we can say that the main objective would be to identify the factors that contribute to the occurrence of accidents. With this information, the competent authorities could plan measures to prevent or reduce such events.

We can conclude that one of the most interested in investigating and verifying the possibility of accident prevention would be the Seattle Department of Transportation.

Data Understanding

The course organizers made the dataset available in csv format. The data is from Seattle Department of Transportation.
The dataset was uploaded on a Jupyter notebook for initial analysis.The main item of the dataset would be the injured people. From this point on, we can verify which other items contribute or not to the occurrence of accidents. Speed, climate, time, day of the week, etc.
We can verify if is there any combination that increases the incidence of accidents. For example, if accidents are concentrated in some region. Which combination of factors could lead to a more serious incident.

Methodology

A profound data analysis was necessary .

At first, list all the columns and understand the characteristics of each one and consider how each one could or could not collaborate for the project. The analysis was carried out by consulting the file “Metadata.pdf” provided by the course.

With this first assessment, it was decided to investigate whether the occurrence of accidents was concentrated in any area. The “Folium” resource was used.

In order to assess which items would be contributing to a more serious accident or not, it was decided by the Machine Learning “Decision Tree” technique.
The reason for using “Decision Tree” would be because we were looking for answers 1 or 2 in the “SEVERITYCODE” column. The “Decision Tree” technique proved to be the most appropriate.

An evaluation was carried out to identify the following actions:

- columns needed or not;
- delete unnecessary columns;
- identify lines with information like “NaN” (Not a Number);
- check the frequency of such NAN information and decide to remove the rows or remove the entire column;
- identify information in the “categorical” format and convert it into a numeric format to enable the application of Machine Learning algorithms;
- check the data balance and, if necessary, apply balancing techniques;
- apply machine learning techniques and verify the accuracy of the responses obtained.

After applying the machine learning algorithm and evaluating the results, some adjustments were made to the algorithm and the data source.

Discussion

We could notice moments when it was necessary to apply functions repeatedly to evaluate various parts of the dataset. Such repetitive work was shortened enormously using Python programming techniques. Throughout the development of the project, we realized the need to observe other details so that the project could be improved, as well as the way of working.

Conclusion

It was possible to perform the prediction and obtain an accuracy index close to 100%. I could notice a possibility to extract more insights from the dataset. More time and study of other treatment techniques are needed, as well as further study of Machine Learning techniques.
We are finishing a stage in learning Data Science, but there is a perception that there is a vast field to be explored in terms of work, as well as, studies for professional improvement.