Coursera Capstone Project

Coursera IBM Data Science Certification

Mohamad Itani

September 2020

Report Content

1. Introduction Section

* The “business problem” to be solved by this project and target audience.

1. Data Section

* Describing the Data Requirements and Data Sources.

1. Methodology Section

* Main component of the report.
* Execute Data Processing, describe/discuss any exploratory data analysis and/or inferential statistical testing performed or machine learning used.

1. Results Section

* Discussion of the results and finding of answer.

1. Discussion Section

* Discussion of observations noted and any recommendations.

1. Conclusion Section

* Answer chosen and conclusions.

1. Introduction
   1. Background Information

Car accidents are the third leading cause of death in the United States with only cancer and heart disease superseding them. As this phenomenon leads to many deaths, we strive to analyze the reasons for its existence and the measures of precautions to be taken.

* 1. Problem at Hand

The goal of this project is to analyze the factors that lead to a higher accident severity in order to take proper measures and mitigate the damage caused. If this project is handled properly and distributed to a large audience, we may ensure the safety of more drivers and lives will be saved.

1. Data Section
   1. Data Requirements

- Accidents data for the United States for the month of June 2020

* 1. Data Sources and Tools Used

- Data set Source: <https://smoosavi.org/datasets/us_accidents>

-Tool: Jupyter Notebook,numpy library,pandas library,matplotlib,seaborn,IBM Watson Studio.

1. Methodology

3.1: Data Preparation:

- In this stage we clean the data of some of its flaws like filling the missing columns with the mean of the column when appropriate and removing the entire column in other situations where it is the more appropriate strategy

- We will also be using the pandas library(for dataframes) along with the numpy library(for the mean) and a cloud based version of jupyter notebook on the IBM Watson Studio to perform our process

3.2: Preliminary Analysis

- The strategy is to begin with plotting accident severity on the y-axis compared with multiple factors to find factors that have a positive correlation with accident severity. This kind of data visualization gives us clues as to where our work must begin.

3.3: Machine Learning

- in this stage we split the data into training and testing sets using cross validation and we train a machine learning model to predict accident severity based on the selected columns as predictors. We use a model (best for our use case is logistic regression because the output is a categorical variable) to predict accident severity for the testing set

3.4 Model Evaluation

- in this stage we use the model evaluation techniques that we have learned to test the accuracy of our logistic regression model and then optimize and rerun the analysis.

4.0 Results

4.1: Leading Factors:

- The leading factor is car accident severity turned out to be weather conditions. This strong correlation shows us the power of driving slowly when the roads are filled with rain as it may be the thing that saves us from our death. The second most influential factor turns out to be road conditions. This is a major signal to government parties to make sure their roads are always in top form. The third most influential condition is light conditions which is obvious because the absence of proper lighting is an almost certain way of increasing a chance of accidents.

4.2: Prediction:

Based on our leading factors of accident severity lead by weather conditions we came up with a logistical regression model capable of predicting accident severity to the 95th percentile approximately. This model is helpful in analyzing accident severity by external factors giving great insight to those who want to learn what makes drivers in the United States get harmed the most.

5.0 Discussion

5.1: What can we do with our findings?

- With the current findings in hand we are able to predict accident severity with a 95% accuracy based on certain factors. This model can be implemented in satellite map applications such as google maps to predict potential accident severity and warn drivers to drive carefully when high severity conditions are met

5.2: Room for improvement

-Further improvements can be made to the model to ensure a higher than 95% chance of predicting accident severity. Moreover, more factors than the one identified in this research can be found making this a more reliable model. As with all creations, there is certainly always room for improvement

* 1. Conclusion

6.1: Key Question:

- The main purpose of this study was to analyze the most influential factors in increased accident severity and those factors turned out to be weather conditions, road conditions, and light conditions. To the best of our ability, we have been able to answer this question.

6.2: Peripheral Questions

-We have been able to predict accident severity with a logistic regression model to a 95% accuracy. This is certainly an achievement in the field of accident prevention. Further improvements are required to perfect the model but this gives businesses a good starting point to base their work on.

Thank You for Your Assistance.