CAR ACCIDENT SEVERITY PREDICTION

APPLIED DATA SCIENCE CAPSTONE PROJECT

ARVIND N

GITHUB LINK: HTTPS://GITHUB.COM/ARVIND-2021/COURSERA_CAPSTONE

Introduction | Business Understanding

Road accidents could be considered an old topic but with the progress of cars and the capabilities of the technology they carry, it is even more important to have models available to predict and mitigate their occurrences. However, predicting the severity of a car crash is not an easy task. Even tough it is possible; precision levels will vary significantly depending on the data available and how well the problem has been modeled.

A machine learning model is required to predict the severity of an accident given the conditions like weather, road and visibility conditions. When conditions are bad, our job is to alert drivers about the increased risk of a car accident.

Description of Data

Data set used for the capstone project is from Seattle DoT. The data comes from records of all collisions provided by SPD and captured in traffic Records. Data includes all types of collisions which occurred at the intersection or mid-block of a segment. Timeframe of the dataset is from 2004 and it's automatically updated on a weekly basis. The dataset comes with a pdf file containing a clear definition for each of the available features.

Our goal is to predict the severity of a crash. In the dataset, the target variable is called 'SEVERITYCODE' because it is used measure the severity of an accident. To achieve our goal, we will go through the following steps:

1. Feature Exploration (with data cleaning)
2. Dimensionality reduction
3. Model building
4. Optimization and final model selection

Considering that our problem is a classification one, we will use an F1 score to evaluate the performance of the different models that we will train. However, we will also take a close look at accuracy, precision and recall since they all provide valuable insights into understanding how the model is performing.

Finally, we will explain and give details on what are some possible next steps to further improve the overall performance and interpretability of the chosen model.