Applied-data-science-capstone (Car accident severity)

1.DEFINITION:
PROJECT OVERVIEW:

Car accidents and crashes are situations we all want to avoid. They are stressful and potentially life-threatening events when we experience fear and adrenaline. If you ever experienced a car accident or a crash, then you probably know how it feels to be frightened for your life.

DATA SET:

The dataset for my capstone project data.govt.nz datasets related to New Zealand and published by the central government from the New Zealand Transport Agency.
It contains data from car crashes on a quarterly basis.

PROJECT STATEMENT AND OUTLINE:

Our goal is to predict the severity of a crash.Where N for non-injury, M for minor, S for serious and F for fatal.
steps:
Feature Exploration
Dimensionality reduction
Model building
Optimization and final model selection

ANALYSIS:
DATA EXLPORATION:

It contains data from car crashes since January 1st, 2000 until the present day and it's automatically updated on a quarterly basis.
The dataset comes with a pdf file containing a clear definition for each of the available features.The dataset has an initial total of 655,697 samples.

Each with 89 different features, both numerical and categorical. Indicating different condition at the time of the crash.

METRICS:

Considering that our problem is a multiclass classification,use an F1 score to evaluate the performance of model to train.Where N for non-injury, M for minor, S for serious and F for fatal.
For each model, the value for the global score will, in turn, be the average of the F1 scores for each class.
So, we will be using a one vs all approach.
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.

MODEL BUILDING:

Benchmarking:
To evaluate the performance of the trained models we will use two reference values.
The first one is the performance of an ANN.
The details are available this paper from 2016 by Sharaf Alkheder; Madhar Taamneh and Salah Taamneh:
Severity Prediction of Traffic Accident Using an Artificial Neural Network
The abstract of the paper describes a final ANN with a test accuracy of 74.6%.

The second reference will be a basic Naïve Bayes classifier used as a baseline for our ensemble models.
For this reference, we will use the averaged F1 scores across all classes as well as the score values at the class level.

DATA PROCESSING:

1.Split the dataset with a 75/25 train/test ratio.
2.Train one baseline model For each variation of the dataset.

The result of this step is one trained model for the original variation; one for the undersampled variation; and one for the oversampled variation.
3.Analyze performance metrics for the baseline models and select the model and variation that performs best.
4.Perform a grid search using the best performing variation.The result of this step is the best estimator based on the averaged F1 score from a 3-fold cross-validation.

RESULT:

The most interesting thing to note is the different approach of two algorithms.
While AdaBoost tries at each step to improve the misclassifications based on a single feature,
Random Forest is a deterministic method that grows trees by finding the split that would produce the greatest reduction of impurity.
Random Forest develops very precise but overfitted trees. however, by adding many trees with a bag of samples for each one, and by bootstrapping the features at each split, it overcomes the overfitting of individual trees, creating a very strong model.

CONCLUSION:

The fact that the overall Accuracy score is only 0.72 indicated that the available data lack,the first benchmark we chose of 74.6% accuracy for the reference ANN.
The Tableau workbook has a few viz to support this conclusion.
Increase of performance as the grid search progresses; indicating that there's only so much the search and models are able to achieve with the available data.
We could also train a binary classifier for each class to have a more customized one-vs-all approach. And also try other evaluation methods like AUC/ROC and Precision/Recall curves.