Coursera Capstone Project : Applied Data Science

Overview

Introduction

- Car crashes and road accidents could be considered an old topic.

Yet, with the progress of the technology involved and the capabilities of these sophisticated machines, it is ever more important to have tools and means available to mitigate their occurrences, as well as their implications and consequences for the people involved.

Introduction

- Thanks to the advancement of technological and analytical tools in the last two decades we are now able to better understand how crashes happen.
- This enables the transport, security and emergency agencies all around the world to have different (predictive) models for quickly analyzing crashes when they happen and dispatch an appropriate response swiftly.

- Many attempts have been taken by many professionals, scholars and government agencies to provide produce these models; each with different goals, ways of measuring success and precision of their results.

Business Problem

- Predicting the severity of a car crash is no easy task. And even when possible, precision levels will vary significantly depending on the data available and how well the system or model has been defined. However, if the dataset's features are clearly defined and if there's a thorough description of how this data is collected –as is our case– we have much better chances of arriving at a usable model. In this particular dataset, the severity for a crash is classified under four different categories: non-injury, minor, serious and fatal. Thus, the main objective of the project will be to create the machine learning model should be able to predict accident "severity.

Business Problem

- This enables the transport, security and emergency agencies all around the world to have different (predictive) models for quickly analyzing crashes when they happen and dispatch an appropriate response swiftly.

Data

The data for this project has been retrieved and processed through multiple sources, giving careful considerations to the accuracy of the methods used.

Data Exploration

The dataset we will be using is an official one published by the NZ Transport Agency. It contains all crashes that have been reported since January 1st, 2000 and it is updated quarterly. The version we will be using includes data up to 2018Q2 included. The dataset has an initial total of 655,697 samples. Each with 89 different features. The feature we want to predict is crashSeverity and it is a categorical feature that takes the values N, M, S and F to indicate that the crash was of class Non-Injury, Minor, Serious or Fatal, respectively.

Data

Categorical Features

The dataset has many categorical features. All these features take a small number of values. Also, they are all nominal features. This means that there's no natural order to their values. Which makes it difficult to transform them in order to be used as input for Sci-kit Learn since it can only handle numerical values. So, as we explain in further detail later on, we will resort to a One Hot Encoding of the dataset –after cleaning it, of course. These categorical features are used to indicate different attributes and condition about the crash. For example, the feature regionDesc correspond to the country's region where the crash happened.

Data

Numerical Features

Most of the numerical features in the dataset are counters that indicate the number of objects involved in the crash. For example, the feature bridge indicates how many times a bridge, tunnel, the abutments or handrails were struck in the crash.

Methodology

Data cleaning

As a first step in the analysis of the data, we have gone through a thorough exploration of each categorical feature and the numerical features that are not counters. The full details of this analysis can be found on sections 2, 3 and 4 of the first notebook. Here, we present the insights and results produced by this analysis. Removal of non-Relevant Features The first things we've done is remove some features that –after considering carefully their meaning as explained by the PDF– we understand are not relevant to our problem. Following is the list of features and the reason why we've decided to remove each one from the dataset.

Feature Distribution

After removing the features mentioned above, we have filled in all missing value with placeholder ###. Then, we explored the values for each of the remaining categorical and numerical features (excluding counters) and treated this placeholder value when necessary.

Methodology

One hot encoding

we tried to uncover some latent features by applying PCA with 10 components to the cleaned dataset after transforming it through One Hot Encoding. We plotted all sampled using the first two PCs which accounted for more than 97% of total variability and saw that the crashes aligned on a few columns (as shown in the viz on the next section). However, these columns concentrate crashes of all classes. Meaning that the first two PCs, which account for more than 97% of total variability, are not able to create a separation between the classes we wanted to predict. Since there were some features with numerical values (the counter) we applied a MinMax scaler to the dataset and repeated the PCA. But results were even worse than before; as the viz in the next section shows.

Methodology

Feature Correlation

Our next step was to explore the correlation between feature. To do this, we computed a Chi2 test of independence for each pair of categorical features. Unfortunately, we weren't ab le to find a pair of features that were independent. In fact, all feature turned out to be dependent on every other feature. This can be seen both in the notebook and the Tableau workbook. A visualization of the dependency of every pair of features is shown in the next section.

Methodology

Algorithms and Techniques

Considering that most of the features in the dataset are categorical and that it is not clear which ones are the most meaningful or relevant, we will focus on decision trees as base/weak learner and train two ensemble models; one for bagging and one for boosting. Specifically, we will train and optimize a Random Forest and an AdaBoost GBM. Random Forest is a great and powerful tool. It leverages the power of decision trees and adds different levels of randomness to overcome their tendency to overfit. This randomness happens when sampling or bootstrapping the instances used to train each tree in the ensemble which makes it a bagging algorithm and again, when a random sample of features are used at each split of each node of each tree. This generates a strong learner able to produce more accurate and stable predictions that any of the weak learners it is made of.

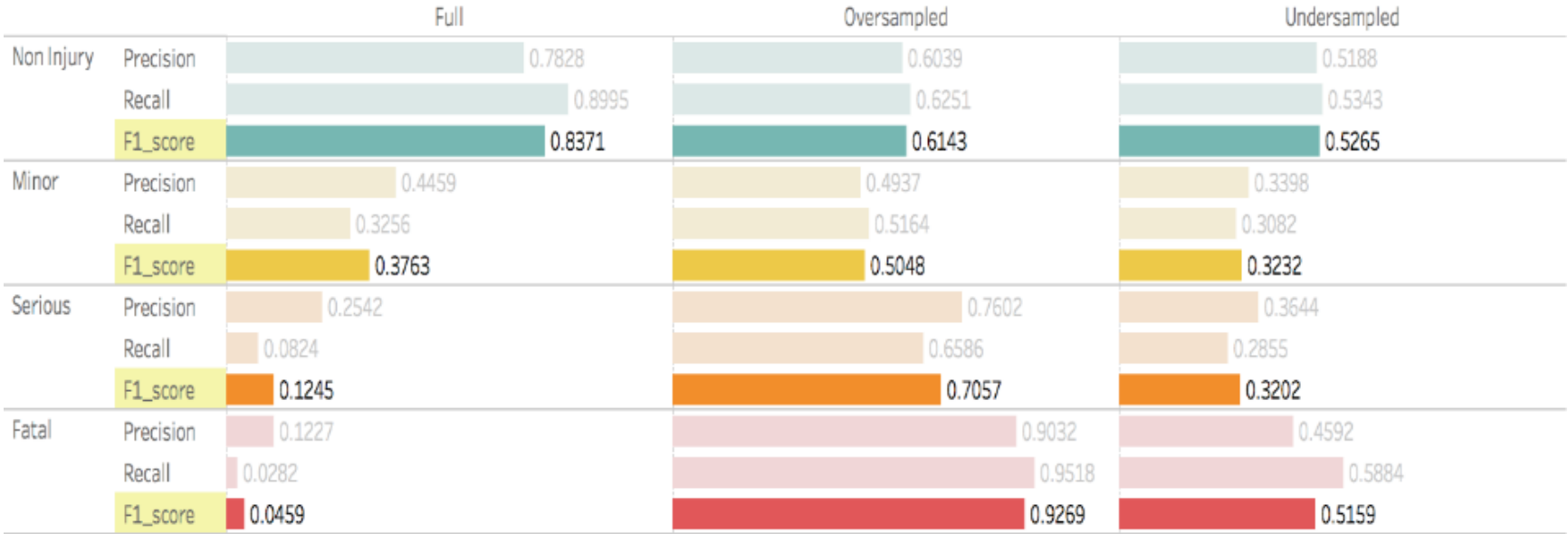
Results

The results obtained by following the process described above were very interesting. In this section we discuss first the performance metrics for the baseline RF models. Then we do the same for the AdaBoost baseline models.

	Accuracy	F1-Score
Original variation	0.7164	0.6861
Undersampled variation	0.4283	0.4207
Oversampled variation	0.6878	0.6879

Results

The in-class performance metrics



Discussion

Given that we have a rich dataset of 655,698 training examples, each with 88 features; and that this data has both a good history going back to the year 2000 and also a good update policy; we have a great opportunity to produce a viable predictive model that could be used by emergency services all around NZ to improve the response time and also to produce a proper response depending on the severity of the car crash. Moreover, given all the features that are available, we have also an opportunity to try to uncover hidden patterns and structure in the data that could be leveraged to better understand which factors are more indicative or play a bigger role in such accidents.

Discussion

Thus, providing valuable insights towards preventing them from happening in the first place (such as better signaling, road maintenance, or even improve road construction planning). Also, using some dimensionality reduction technique such as PCA, Random Projection or ICA could provide us with new and unseen features while also simplifying dataset used to train the classifiers.

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
60	Bijpur, North 24 Parganas	Train Station	Women's Store	Electronics Store	Film Studio	Field	Fast Food Restaurant	Falafel Restaurant	Fabric Shop	Event Service	Eastern European Restaurant
123	Garshyamnagar	Train Station	Platform	Women's Store	Eastern European Restaurant	Field	Fast Food Restaurant	Falafel Restaurant	Fabric Shop	Event Service	Electronics Store
131	Halisahar	Train Station	Women's Store	Electronics Store	Film Studio	Field	Fast Food Restaurant	Falafel Restaurant	Fabric Shop	Event Service	Eastern European Restaurant
141	Hind Motor	Light Rail Station	Train Station	Women's Store	Electronics Store	Field	Fast Food Restaurant	Falafel Restaurant	Fabric Shop	Event Service	Eastern European Restaurant
186	Kodalia	Train Station	Women's Store	Electronics Store	Film Studio	Field	Fast Food Restaurant	Falafel Restaurant	Fabric Shop	Event Service	Eastern European Restaurant

Figure:Cluster having Train Station as most common venue

Conclusion

- The initial goal of this project was to train a model that would be able to predict the severity. While working on this problem and dataset I learned many things about working with categorical data, multiclass classification and class imbalanced. I has to explore deeply and thoroughly these features to make sure I wasn't introducing any bias into the model. I had to remove featured because they produced data leakage and had to treat both missing and inconsistent values for different features. Understanding the relevance of the different features was particularly hard for me because of their categorical nature. The results obtained by applying PCA actually provided more question rather than answers.