Road Safety Mini Report

How likely is an officer to be present at the scene of an accident?

by
Christian Lopez

First, I will answer some of the questions given in the assignment followed by what I learned in this experiment. Some assumptions made for this assignment are that the data has been slightly changed from the original dataset (e.g. the feature *Did_Police_Officer_Attend_Scene_of_Accident* shows 3 categories in the data guide; however, only 2 of those are present in the given dataset) and that I am restricted to using **DfTRoadSafety_Accidents_2014.csv** only.

a. Describe the features and the target included in the dataset. What insight did you gain from working with the data?

The most notable issue with the dataset and in particular with the target feature is that it is a very imbalanced dataset with plenty of examples of when police is present in an accident and very few of no police presence:

Present	Not Present
119607	26715

The biggest risk of using the data unchanged is the introduction of bias and inaccurate performance of the model. To address this issue Synthetic Minority Oversampling TEchnique (SMOTE) is applied to the dataset in an attempt to improve the model's accuracy—note: f1-score without SMOTE capped off at 0.51 while the SMOTE set provides an f1-score of 0.75).

b. Explain how you transformed the features

The first transformation performed is to the 'Time' feature, this column is formatted using a 24hr time interval including minutes, changing this feature from an object type to a integer type including only the hours might help the model perform better. Further, this category is considered to be more helpful if it can be encoded into different categories—such as morning commute hours and late afternoon rushes—the reason that it is believed that this will be helpful is that accidents are more likely to occur when more vehicles are present, and this might incur more police presence. This transformation proved to be helpful when analyzing the most important features of the model.

Next, certain features miscategorised as 'objects' or 'ints' are casted as categorical data—their correct datatype—this is done to help create a more accurate model. Further, all of the categorical data is separated into its own column using one-hot encoding, this is done to avoid confusion by the model which will lead to a more accurate prediction.

c. The validation approach taken

Cross-validation is for model validation. Through the research done for this project I found that cross validation is well suited for imbalanced dataset.

d. Which performance metrics did you use and why?

Some of the performance metrics used in the model are

Confusion Matrix:

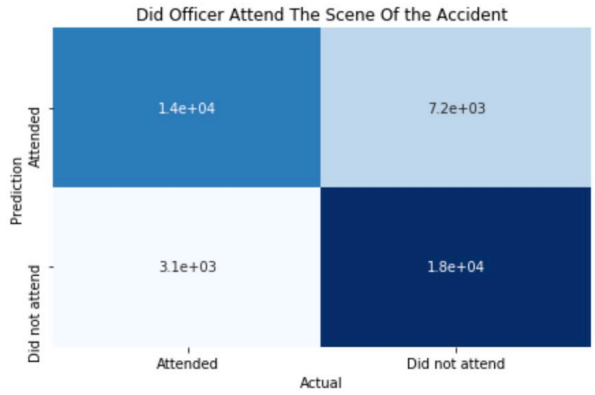

Fig1: Police presence confusion Matrix

Confusion matrix was chosen to visually inspect the categories falsely/correctly classified by our model. Further, the Root Mean Square Error (RMSE) was used to evaluate the model in order to find the absolute fit of the model the given data.

```
Classification Report Random Forest - with Entropy and SMOTE Upsampling:
               precision    recall  f1-score   support

           1       0.81      0.65      0.72     20754
           2       0.71      0.85      0.78     21021

    accuracy                           0.75     41775
   macro avg       0.76      0.75      0.75     41775
weighted avg       0.76      0.75      0.75     41775

RMSE 0.4983995930963063
Cross Validation

array([0.74339982, 0.74138047, 0.73790635])
```

e. Which algorithms did you use?

In the model submitted I used a Random forest algorithm as it is a well establish and widely adapted supervised learning model for classification, I believed that it might be a good base to test different models against in future experiments. I also ran more experiments using XGboost algorithm and finally tried using H2O open source auto ML for comparison.

f. Which were the most important features?

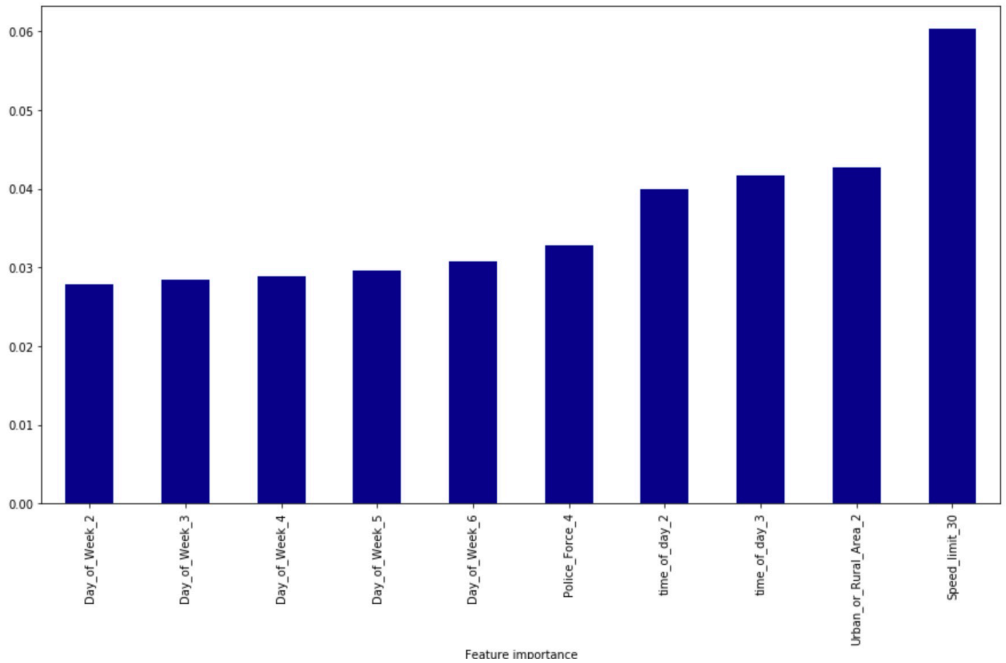

Fig2: Most important features

Some of the most important features are the time of day that the accident occurred and whether the accident took place in a rural or urban area. I expected that the time of day that the accident occurs has a big influence in police presence and this is confirmed by the model; however, I did not the 'Urban_or_Rural_Area' feature to have a big impact in our results.

g. How useful is the model from a practical point of view?

I believe that in order to use this model in a practical situation we must improve the accuracy by providing more data. Further, more validation is needed to further ensure the performance of the model.

h. What might you do differently if you had more time/resource?

If I had more time available, I would like to run more experiments to try different algorithms and validation techniques. Most importantly, I would of liked using the entire 'Road-Accident-Safety-Dataset' while reading the guide of the dataset I noticed that some features are not included in DfTRoadSafety_Accidents_2014.csv and some insight might be lost because of it.

Discussion:

For comparison against the random forest algorithm H2O open source and XGboost are used. These are the results provided by H2O Auto-ML

model_id	mean_residual_deviance	rmse	mse	mae	rmsle
StackedEnsemble_AllModels_AutoML_20190703_172148	0.138022	0.371513	0.138022	0.278553	0.151298
StackedEnsemble_BestOfFamily_AutoML_20190703_172148	0.138052	0.371553	0.138052	0.278659	0.151318
GBM_5_AutoML_20190703_172148	0.138563	0.372241	0.138563	0.277906	0.151494
XGBoost_1_AutoML_20190703_172148	0.138695	0.372418	0.138695	0.276878	0.151497
GLM_grid_1_AutoML_20190703_172148_model_1	0.138716	0.372446	0.138716	0.282321	0.151756
GBM_3_AutoML_20190703_172148	0.138896	0.372687	0.138896	0.279098	0.15163

The XGboost model yielded a similar RMSE value as the StackedEnsemble_AllModels_AutoML_20190703_172148 at RMSE: 0.372263

Reference:

https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

https://medium.com/lumiata/cross-validation-for-imbalanced-datasets-9d203ba47e8

https://www.marcoaltini.com/blog/dealing-with-imbalanced-data-undersampling-oversampling-and-proper-cross-validation