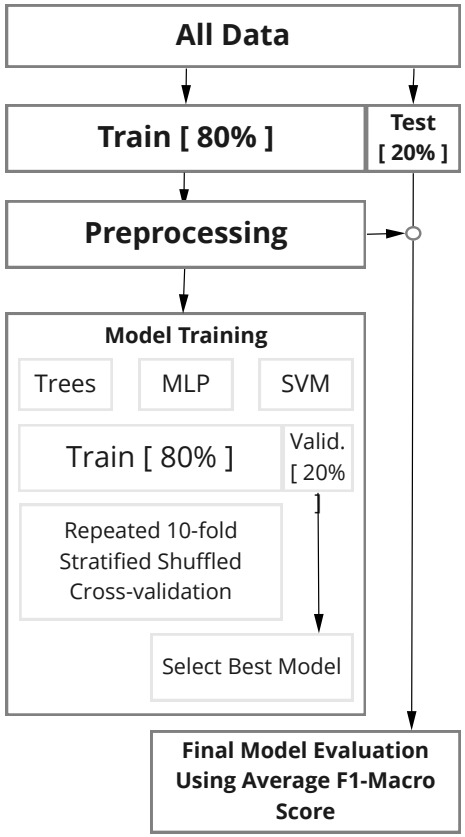
Analyzing New York Police Racial Profiling With A Machine Learning Approach

Alex Martorell & Louis Van Langendonck - Machine Learning Course 2022 Final project

Abstract

The **goal of this project** is to try and **predict race** based on the "Stop, Question and Frisk" data set of 2020. This data describes the working procedures of The New York Police Departement. The resulting models can help provide insight into the potentially present racial profiling.

Pipeline Overview



Preprocessing

Data set description: General things covered are the location of the stop (Street, borough, police sector...), the suspect description (Age, height, weight, race, demeanor...) and some general binary stop specifics (Physical force used, weapon found..). In total, there are 9544 rows and 83 columns (72 categorical, 10 numerical and 1 time stamp variable). The distribution of the target variable is imbalanced as can be seen in Figure 1.

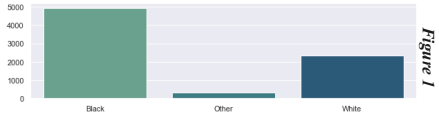


Figure 1

In **data cleaning** the data set is explored, adequate data types set and naming conventions adopted. Gaining familiarity with the category-heavy data is achieved by plotting cross-tabs. For example, in Figure 2, the distribution of boroughs is plotted in function of if the suspect is arrested or not.

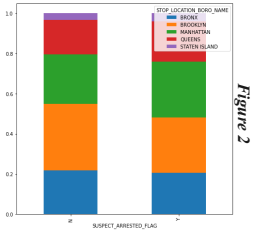


Figure 2

Because of heterogenity in missing value notation and data mistakes, **missing values** are identified column by column. Columns and rows containing excessive amounts are omitted. KNN imputation is applied on the suspects Age, Height and Weight variables.

Regarding **outliers**, two approaches are maintained. For categorical variables, minority levels are collapsed. For numericals, both stop *duration* and observed *duration* contain outliers. The former is rescaled while the latter is converted into a binary predictor.

In **variable derivation**, the suspect demeanor is looked at, which describes in words the suspects behaviour. As it contains 7570 unique values, it can be considered natural language data. To be able to use is as a predictor, a simple NLP pipeline is build and the text clustered in 11 distinct groups.

For **feature selection**, statistical tests (Chi2 and Kendall) are applied as a first filter. The extra-trees classifier adds a second layer to obtain the final number of predictors.

MCA and Clustering

For **Clustering,** an all-numerical data set is needed. Therefore **MCA** is applied on categorical values. The first 10 dimensions explain about 25% of the variance. The first dimension separates arrested and not arrested suspects. For clustering both *K-means* and *EM* are used. Both group the data into two clusters as an optimal solution.

Models

Tree-based Models

Different tree families are tried: *Decision Trees, Random Forests and Extra-Trees*. For each of them, parameters are optimized using grid-search according to cross-validation described in the pipeline overview. The best performing model is a Random Forest with parameters: *bootstrap = False, Class Weight = balanced, Max Depth = 20, Minimum Samples leaf = 4, Minimum Samples split = 2, Number of Estimators = 200.*

Multilayer Perceptron

For the Multilayer Perceptron model, the same grid-search strategy applied. The best performing architecture is a single-layer containing 9 neurons with the *logistic* activation function, the '*adam*' optimizer and an *alpha*-regulator of value 1e-05.

Support Vector Machine

Different kernel functions are tried: *linear, polynomial, radial basis* and *sigmoid*. Note that SVMs have are slow in their training, hence a less extensive grid-search is applied. It yields an optimal *polynomial kernel* with the following parameters:
C = 0.1, y = 10.0, d = 1, r = 1

Result Overview:

	Avg. Macro F1 (%)	Per-class Accuracy (%) [Black, Other, White]
Random Forest	50.7%	[78.4, 19.2, 54.3]
Multilayer Perceptron	54.7%	[82.3, 23.4, 58.5]
Polynomial Kernel SVM	43.0%	[70.9, 14.0, 45.8]

Results

With the MLP as the best performer on the validation set (F1 = 54.7%), it is chosen as final model. On the **test set**, it yields a **Macro F1-score of 53.4%.** Given how similar the test- and validations socres are, it can be concluded that the model has a **low generalization error.** For a stop, it predicts if it was a Black person with 81.7 % accuracy, a White person with 55.6 % accuracy and a person from different race with 23.0 % accuracy. The confusion matrix is shown in Figure 3.

Figure 3

In order to gain insight in what are the best predictors, the best Random Forest models feature importance is used. The top 11 predictors are shown in Figure 4. It can be deducted that:

- **Black people** stopped by the police are **generally younger**, which makes for the best predictor.
- **Stops of Black people** are generally **shorter** and happen **later in the day**.
- Most of the **'Others' group** are predicted by the fact of the stop happening **in Queens.**
- The police describes the **behaviour of Black people** stopped often as either **not calm or cooperative**, while the **'Other' group** often gets described as either **calm or non-cooperative.**

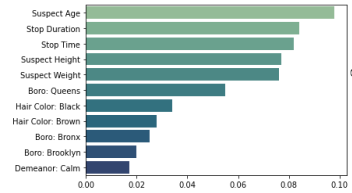


Figure 4

Conclusion

Race is predicted suffciently accurate (F1: 53.4%) considering the complex nature of the concept. The highest contribution variables are *Age, Stop Duration* and *Stop TIme.*