Data Study - Animal Rescues in London

Stéphane Branly and Quoc Hung Tran

3 mars 2023

Résumé

Within the framework of the SY09 course (Data Science), a project takes place with the aim of applying the notions seen in class on a dataset. The selected dataset is the one of Animal Rescues in London proposed during the data analysis challenges #**TidyTuesday**. This paper shows an analysis of the dataset as well as the intellectual path we took to perform this analysis. This analysis is guided by set objectives. The whole code is available on the Github repository SY09-Project

1 Introduction

1.1 Introduction to the dataset

The dataset contains 31 variables and is available as a .csv file. This dataset contains all the animal interventions (animal rescues) including location information (postal code, district, neighborhood, GPS coordinates), date and time of the interventions as well as details on the intervention (reason, call location, animal, cost). These are the interventions recorded in London between January 2009 and May 2021.

1.2 Study objectives

Given the type of data : space-time series, we decided to understand the involvement of spatial and temporal aspects in interventions. To explore the data, we will thus focus on two main objectives.

1.3 The rescued animals

Our first objective will be to understand which animals are rescued during interventions. More specifically, we want to understand the relationship between the type of animal and the location of the intervention, the date of the intervention, the blow or the type of intervention (ìrescued animal at height, in the water, ...î).

1.3.1 The types of intervention

This second objective will allow us to understand what the types of interventions are and especially how they depend on the location, the date, ...

1.4 Exploring the data

1.5 Cleaning the data

Before starting to analyze the data, a first phase of data cleaning is necessary in order to make them usable. The 31 variables were thus analyzed and cleaned. The cleaning consisted in completing the missing location information (*latitude, longitude*) thanks to the other columns and then removing the columns duplicating the location information in different forms. The correlation matrix also allowed us to remove trivial linear relationship columns that we were not going to use for the rest of the study (*hourly_notional_cost* and *pump_notional_cost*).

For the rest of the study, we work with a dataset containing 16 variables which are the following :

Table 1 – Dataset variables

date_time_of_call	animal_group_parent
pump_count	incident_notional_cost
originof_call	borough
property_type	property_category
special_service_type_category	special_service_type
latitude	longitude
month	year
dayofweek	hour

1.6 Analyse des données

1.7 Spatialité

Next, we were interested in the spatiality of the interventions according to the animals. With the help of

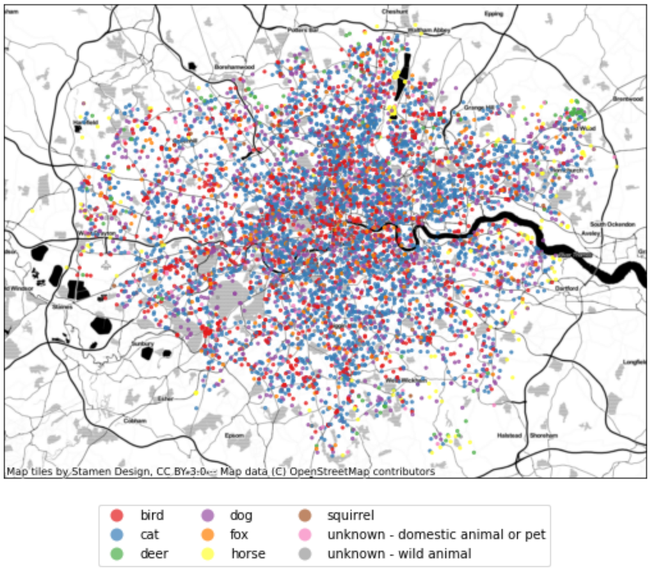
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors
bird
cat
deer
dog
fox
horse
squirrel
unknown - domestic animal or pet
unknown - wild animal

FIGURE 1 – Geolocation of interventions

a map (1 as well as the one available in appendices (9)), we notice that the interventions of domestic animals or of small sizes (*cats, dogs, birds*) are rather concentrated in the center of London. While interventions on rural/wild animals or large animals (horses, deer) are concentrated on the outskirts of the city.

1.8 Temporality

We also decided to study the temporality of the dataset and in particular to see the differences in occurrences of intervention according to the types of animals. To do this we looked at different temporal scales. We focused on the centers of inertia per group of animals. For example, we can see in the figure that squirrel interventions occur mainly in July and August while dog interventions are evenly distributed over the different months of the year. Center of inertia analyses were also done for day of the week and time of day.

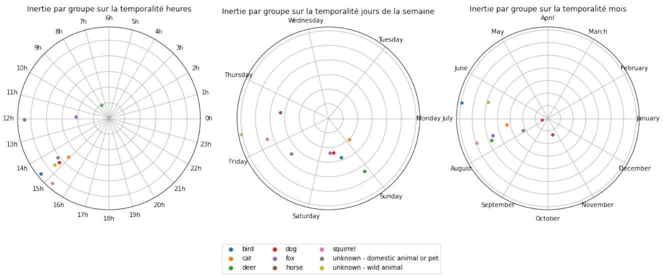

FIGURE 2 – Temporal center of inertia by animal type

1.9 Principal component analysis

After cleaning the data, we performed a principal component analysis to consider the reduction in the number of explanatory variables in order to allow visualization of the data given their high dimensionality. We notice that the axis PCA_1 explains 27.6 % of the total variance and PCA_2 22.9 %. Thus, with these two components we can represent 50.5% of the information contained by the 16 variables. This is a low percentage to hope to perform data compression. Especially since the shape of the cumulative inertia is linear rather than logarithmic.

The visualization of the data on the two main axes of the PCA can be seen in figure 3.

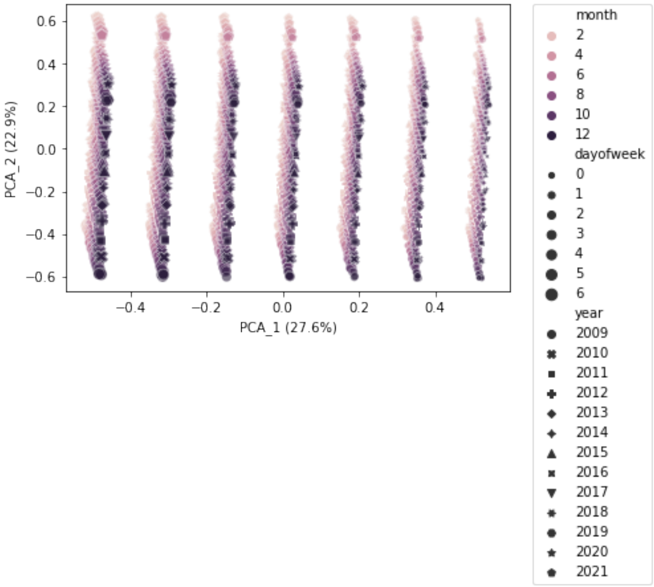

FIGURE 3 – Representation of the data set on the 2 main PCA axes

Remarks : Here, we have axes mainly representing temporal variables. It is interesting to see that spatiality and intervention cost are not distinguished. Ideally, we would have wanted a PCA that would allow us to distinguish the animal or type of intervention, but unfortunately, we do not distinguish any groups for these qualitative variables.

1.10 K-means partitioning

We also decided to do a k-means partitioning on our data set. Unfortunately, the low number of numerical variables means that the classes created do not allow us to associate them with the corresponding classes for *type of intervention* or *type of animal*. The same is true for our *improved* data set, whose construction is detailed in section 2.2.1, but we will notice the impact of adding variables related to spatiality. The visualization of the k-means partitioning is visible in appendix 4.

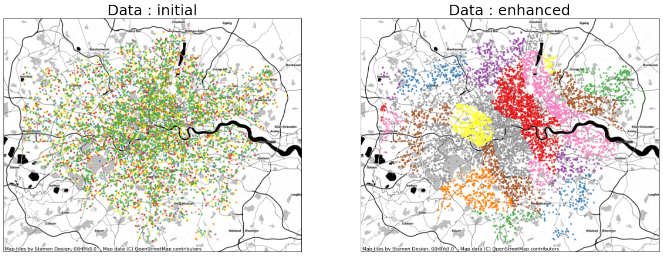
Data : initial
Data : enhanced
Map tiles by Stamen Design, CC BY 3.0 — Map data (C) OpenStreetMap contributors

FIGURE 4 – Representation of classes resulting from the k-means partitioning method

1.11 Hierarchical Ascending Classification

Finally, we also performed a hierarchical classification by grouping data sets by *animal type* and *type of intervention*. The numerical data used corresponds to the averages. It is interesting to see that the *improved* data set brings *horses, cows, and goats* together, as well as the *cat* with the *fox*. But also that *height and ground-level interventions* are distant from *aquatic interventions*.

2 Supervised Classification

2.1 Introduction

After data analysis, we present the different methods applied to solve the classification problem.

Here, the objective of the classification is to predict the **type of intervention**, then the **type of animal**, which we mark as Y. The variable to be explained, y, is nominal qualitative with 4 modalities : *Animal rescue from water, Animal rescue from height, Animal rescue from below ground, Other animal assistance* for the type of intervention. And with 10 modalities : *cat, bird, dog, fox, horse, unknown - domestic animal or pet, deer, unknown - wild animal, squirrel, unknown - heavy livestock animal* for the type of animal.

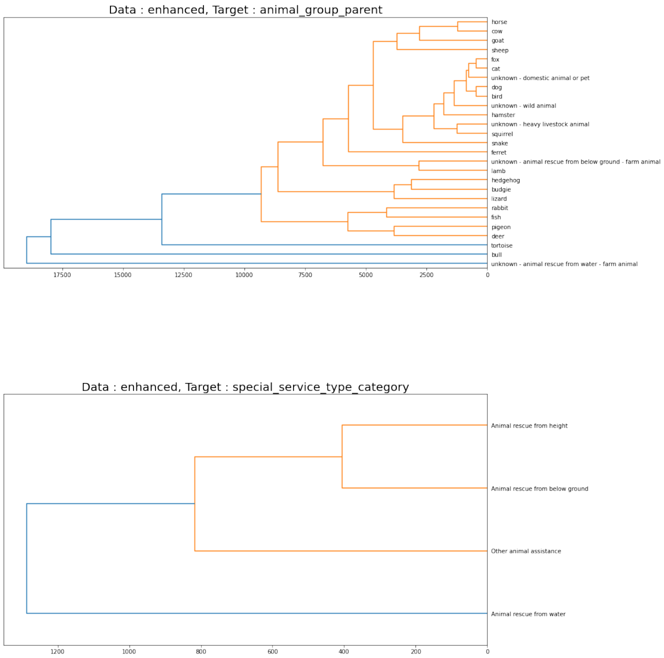

FIGURE 5 – Dendrograms resulting from the ascending hierarchical classifications

2.2 Preparing data for classification algorithms

2.2.1 Expertise Contribution

We were able to apply some supervised classification methods, but the results were not satisfactory. We therefore decided to **bring expertise in the field**, particularly through our experience, reading articles related to animal interventions in London, and analyzing the dataset. We added geographical information by indicating the distance from the nearest physical objects (*forest, fields, lakes, commercial areas, residential areas, etc.*) for each intervention. This adds quantitative variables that make sense in relation to our targets (*type of animal and type of intervention*). The geographical data used comes from OpenStreetMap and allows us to retrieve *features* by geometry type (*Point, Line, Multiline, Polygon, etc.*) and by *tags*.

The data was retrieved for the city of London (represented in Figure 6).

We also decided to perform transformations on the temporal variables. Indeed, the order of the *month*, *day*

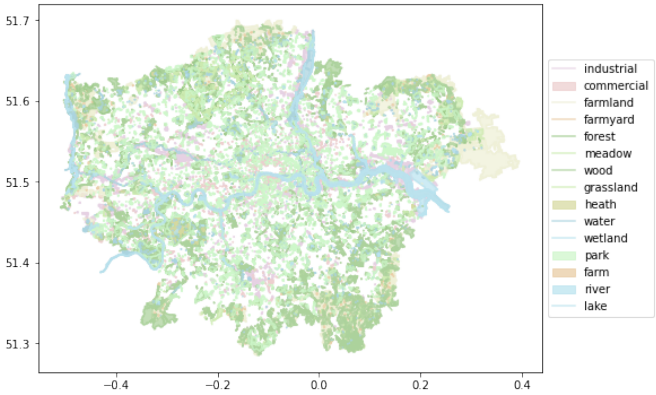

FIGURE 6 – Representation of geographical data retrieved for London

of the week, or even *hour* meant that, for example, the distance from *January to December*, from *Monday to Sunday*, or from *00h to 23h* was greater than the distances respectively from *January to February*, from *Monday to Tuesday*, or from *00h to 01h*. These variables were then transformed into *cyclic form on 2 variables* in order to avoid ordering problems. Figure 2 illustrates the transformation performed.

The dataset using this expertise and transformation will be called the **enhanced dataset**, while the initial dataset will be called the **initial dataset**.

2.2.2 Data preprocessing

The first step in any machine learning project is to transform the data into a format that can be used by machine learning algorithms. In most cases, we need numerical data without missing values or outliers that could make it much more difficult to learn a model.

In the case of the dataset, the values are already cleaned up, but we will need to transform the data into numerical values.

2.2.3 One-Hot Encoding

For categorical variables, we performed transformations using One-Hot Encoding. This creates as many Boolean variables as there are modalities for our categorical variable. The variable is set to *True* if it corresponds to the modality, and *False* otherwise.

2.2.4 Dimensionality reduction, variable selection

We also decided to add a dataset with reduced dimensions using a variable selection method. We used our enhanced dataset that contains the most columns. The variable selection method used is recursive feature elimination, which allows us to establish an importance ranking for each variable. This ranking is calculated using an estimator. Here we used the logistic regression estimator. Variable selection allows us to halve the number of variables. We will see later if this affects the results of the algorithms.

The dataset containing the selected variables will be called **selected features**.

2.2.5 Partitioning datasets

Our datasets were partitioned to find hyperparameters, train and compare methods with cross-validation, and test results. Thus, we shuffled the datasets (changing the order of the individuals) to avoid retaining the temporal order, then we split the datasets into three parts. We take care to keep the same random seed to shuffle individuals from different datasets (*initial, enhanced, and selected features*).

2.3 Methods used

We used several supervised classification methods for our datasets and targets.

2.3.1 Hyperparameters

Some of these methods require finding hyperparameters such as the K-nearest neighbors method or Random Forest.

For example, for the K-nearest neighbors algorithm, to determine the optimal number K of neighbors, we used a **grid search**, which is an optimization method that allows us to test a series of parameters and compare performance to deduce the best parameterization. There are several ways to test the parameters of a model, and grid search is one of the simplest methods. We then verify that the K found is consistent with another data sample.

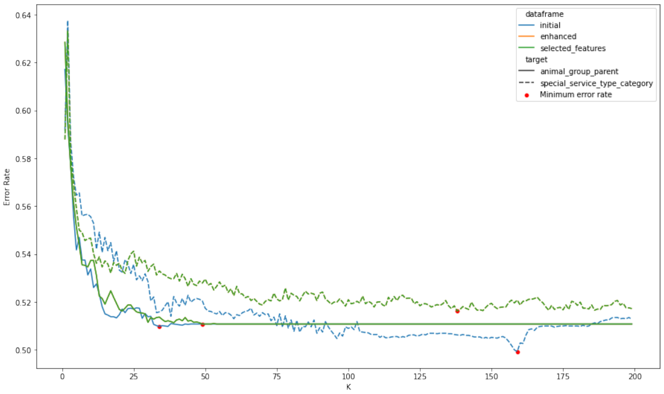

Figure 7 – VVisualization of the error rates for the samples as a function of K

2.3.2 Comparison of Models and Dataset Versions

Once the hyperparameters have been found, we can use our training datasets to compare the different models with cross-validation. At this stage, we take care to set the random seed of the cross-validation so that it is similar from one method to another.

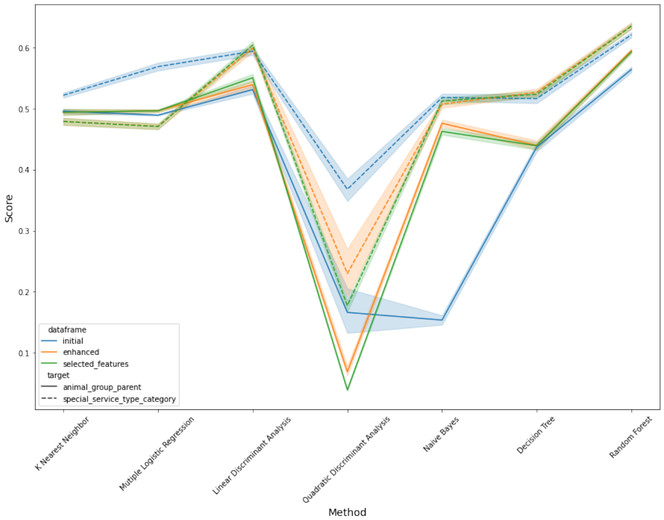

Figure 8 – Comparison of model scores

Several observations can be made when comparing the different models and dataset versions.

First, we notice that the *improved* and *selected variables* datasets tend to give better results than the *initial* dataset. This is a good sign, the geographic and temporal expertise added allows the models to classify better. This also means that our variable selection did not deteriorate the *improved* dataset. On the contrary, it even improves the score. In the context of climate change, it is wise to note that we have reduced the number of variables by half while maintaining the quality of classification.

Secondly, we can see that the *intervention type* target generally obtains a better score compared to the *animal type* target. This can already be explained by the different number of categories (4 and 10 categories, respectively).

Finally, we can note that the Random Forest and Linear Discriminant Analysis methods provide the best scores for both targets. The decision tree already has a decent score, but the forest of trees, which contains multiple decision trees, boosts its performance. The performance of the linear discriminant analysis model can be justified by the normal distribution of classes as well as the fact that the classes are similar in terms of orientation and volume. The quadratic analysis has a much lower score because it only assumes a normal distribution, and therefore the number of parameters to find is much larger.

3 Conclusion

In conclusion, we have seen some analyses of the dataset : spatial and temporal visualizations, PCA, hierarchical clustering, k-means... We were able to distinguish some classes with these methods, but we saw that they did not correspond to our classification expectations.

On the other hand, supervised classification allowed us to create and train a classifier on the desired target. We were able to compare different models, datasets with the addition of expertise and variable selection. We then obtained good results that we tried to justify based on the classifiers.

We also invite you to see our code on the Github repository SY09-Projet to see all the operations performed on this dataset.

4 Further work

We have decided not to delve further into some points that have been mentioned in this report for several reasons : human resources to allocate, lack of expertise in the field, outside the scope of UV and SY09 project.

For example, we can mention the expertise that was brought in, which is interesting but could have been exploited even more. We could have indicated the number of elements present within x meters, indicated the area of polygonal elements, added the length of line elements... Furthermore, for more rigor, it would have been necessary to take into account the elements also present around London (and not just in London), to pay attention to the dates of existence of these elements compared to the dates of interventions. We could also have provided other information such as the weather on the day of the intervention, the brightness (day/night)...

It would also have been judicious to see if we could perform mathematical transformations on certain variables, such as applying the exponential function, square...

The targets used here do not have practical meaning if we wanted to create a tool to assist the London fire brigades. For example, we could have removed variables determined after an intervention (cost of the intervention, for example) and then created a tool to predict the location of the intervention and the type of animal and intervention.

There are indeed studies on predicting the location of interventions (in the case of crimes, for example) that we could exploit. One of the implementations is based on seismic models. We invite you to watch a popularization video on the subject : Can we predict future crimes? Crowd counting.

Finally, to complement the point that has just been mentioned, we could also have looked at classification models that take into account previous operations in predicting new ones.

5 Annexes

5.1 Figures

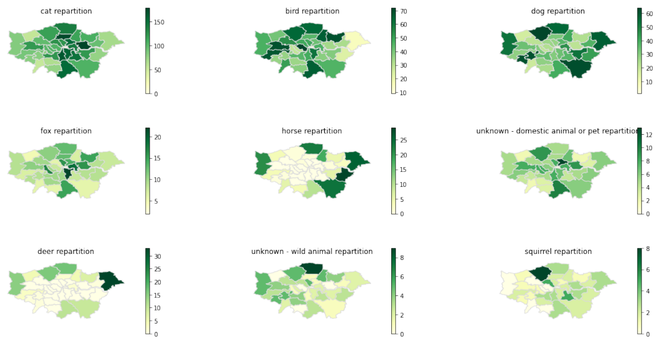

FIGURE 9 – Distribution of interventions on districts by type of animal

Références

[1] *The London Fire Brigade rescues hundreds of creatures every year – from pigs to cats to bearded dragons.*, published on the London Fire Brigade website.

[2] Perkin Amalaraj,

[3] Constance Kampfner,

[4] *London's firefighters attended two animal rescues a day in 2020*, published on January 7, 2021 on the London Fire Brigade website.

[5] M. B. Short, M. R. D'Orsogna, V. B. Pasour, G. E. Tita, P. J. Brantiham, A. L. Bertozzi and L. B. Chayes, *A statistical model of criminal behavior*, published on December 28, 2007.