UNIL | Université de Lausanne

Machine Learning Final Project

Machine learning in Business Analytics 2024, Group C

Lodrik ADAM — Student id: 18418871
Stefan FAVRE — Student id: 18826333
Jeff MACARAEG — Student id: 18876233

2024-05-06

Table of contents

Introduction • The context and background: course, company name, business context.

During our 1st master year as students in Management - orientation Business Analytics, we have had the opportunity to attend some lectures of Machine Learning for Business Analytics. In content of this class, we have seen multiple machine learning techniques for business context, mainly covering supervised (regressions, trees, support vector machine, neural networks) and unsupervised methods (clustering, PCA, FAMD, Auto-Encoder) but also other topics such as data splitting, ensemble methods and metrics.

• Aim of the investigation: major terms should be defined, the question of research (more generally the issue), why it is of interest and relevant in that context.

In the context of this class, me and my group have had the opportunity to work on an applied project. From scratch, we had to look for some potential dataset for using on real cases what we have learned in class. Thus, we had found an interesting dataset concerning vehicule MPG, range, engine stats and more, for more than 100 brands. The goal of our research was to predict the make of the car according to its characteristics (Consumption, range, fuel type, …) thanks to a model that we would have trained (using RF, ANN or Trees). As some cars could have several identical characteristics, but could differentiate on various other ones, we thought that it would be pertinent to have a model that was able to predict a car brand, from its features.

• Description of the data and the general material provided and how it was made available (and/or collected, if it is relevant). Only in broad terms however, the data will be further described in a following section. Typically, the origin/source of the data (the company, webpage, etc.), the type of files (Excel files, etc.), and what it contains in broad terms (e.g. "a file containing weekly sales with the factors of interest including in particular the promotion characteristics").

The csv dataset has been found on data.world, a data catalog platform that gather various open access datasets online. The file contains more than 45'000 rows and 26 columns, each colomn concerning one feature (such as the year of the brand, the model, the consumption per barrel, the highway mpg per fuel type and so on).

• The method that is used, in broad terms, no details needed at this point. E.g. "Model based machine learning will help us quantifying the important factors on the sales".

Among these columns, we have had to find a machine learning model that could help us to quantify the importance of the features in predicting the make of the car. (To do so, we tried to use ANN model and RF)…

• An outlook: a short paragraph indicating from now what will be treated in each following sections/chapters. E.g. "in Section 3, we describe the data. Section 4 is dedicated to the presentation of the text mining methods…" In the following sections, you will find 1st the description in the data, then in Section 2 the method used, in Section 3 the results, in Section 4 our conclusion and recommendations and finally in Section 5 our references.

Data description

• Description of the data file format (xlsx, csv, text, video, etc.)

On the webpage of the dataset, we have found a CSV and xlsx format. We have decided then to download both.

• The features or variables: type, units, the range (e.g. the time, numerical, in weeks from January 1, 2012 to December 31, 2015), their coding (numerical, the levels for categorical, etc.), etc.

In the original dataset, we had the following 26 columns, each one corresponding to a a feature. Here is a quick overview of the types of variables. (MAKE TABLE)

• The instances: customers, company, products, subjects, etc.

In a basic instance, each row is concerning one car. We can find in order the ID of the car corresponding to a precise feature observation, then the features as seen in the table before.

• Missing data pattern: if there are missing data, if they are specific to some features, etc.

• Any modification to the initial data: aggregation, imputation in replacement of missing data, recoding of levels, etc.

• If only a subset was used, it should be mentioned and explained; e.g. inclusion criteria. Note that if inclusion criteria do not exist and the inclusion was an arbitrary choice, it should be stated as such. One should not try to invent unreal justifications.