Capstone proposal for obtaining the Udacity Machine Learning Engineer Nanodegree

The proposed capstone project will deal with customers data coming from Arvato Financial Solutions, a Bertelsmann subsidiary. Arvato is a global services company including customer support, IT, logistics and finance.

Arvato requires that, with the provided dataset, one provides a model able to state whether or not a person will respond to a specific campaign and become a customer. This requires being able to perform customers segmentation i.e. devising customers clusters based on several features, identifying the key features that distinguish a customer from a non-customer and eventually building a model using the identified features to state whether or not a given person will (or how likely he is to) become a customer.

Arvato provides the dataset being composed of four different files :

- A general population file containing 891 211 samples (persons) with 366 features for each sample;
- A customer file containing 191652 samples with 369 features i.e. 366 features are identical to the general population file while three additional ones are provided : CUSTOMER_GROUP, ONLINE_PURCHASE and PRODUCT_GROUP identifying the type of customer being described;
- A TRAIN mail campaign file containing 42982 samples and 367 features i.e. 366 features being identical to the general population file and one additional feature identifying whether or not the person became a customer following the campaign;
- A TEST mail campaign file containing 42833 samples being identical to the TRAIN file except that the target (became customer) has been removed.

The problem at hand is therefore a supervised leaning problem. One needs to use the general population file and customer file to try devising the features that distinguish a customer from a non-customer. Once these features are identified, one will build a model to infer if a given person will become a customer or the likelihood of this event for that particular person.

To solve this problem, the following techniques will probably be used :

- Dataset exploration using visualizing techniques such as pair-wise plot or grid plots;
- Feature-space reduction technique e.g. Principal Component Analysis;
- Supervised model such as XGBoost, SVM or a neural network

The supervised model shall be trained using probably a binary cross-entropy loss while the model shall be evaluated using an accuracy metric based on the training & test datasets provided. The accuracy will determine how the model fares at determining the right class for a given person i.e. will he become a customer or not based on that person's features.

The output for this project will consist in Jupyter Notebooks outlining the different steps described hereabove in details. The Jupyter Notebooks will contain all code, models, data analysis and graphs necessary for the study and will be written in Python. The Jupyter Notebooks will be elaborated under the AWS SageMaker environment and will therefore require access to this particular tool to be executed as the code will call for aws ressources. If necessary, model artifacts could be directly provided under a zip file.

BULLE Jérémy