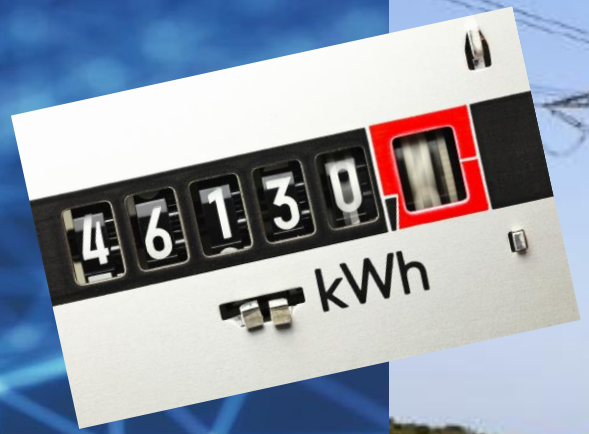


PREDICTING ELECTRICITY CONSUMPTION

PIERRE CHAVANNE – 03/10/2023

PROJECT OVERVIEW

- A national electricity producer addressed our consultancy firm a project about future electricity demand prediction.

- Raw statistics are available after a huge collect of consumers data from all over this country :
  - Area spotted : Nothern, Southern, Western, Eastern, Centre.
  - Parameters : fall-winter and spring-summer average temperatures, electricity demand and energy independancy rate (0 if consumer uses 100% electricity as source of energy, else if other power sources used as gas, domestic installations…).

- Goals :
  - Identify groups of consumers after this dataset (representative of the market)
  - Predict future electricity demand after simple public features as temperatures : this dataset is complete but expensive to collect and thus too difficult to get in the future.

DISCLAIMER

- This dataset has been created by my own and is not representative of a real use case :
  - Only 500 rows of data to compute quickly
  - Only 5 features
  - Created after a manual mix of all features : displayed data can seem a little bit « in boxes »

- This is more a proof-of-concept of ML algos in clustering – multiclassification - regression.

ROADMAP

- 1. Analyze dataset to detect two features over which groups of electricity demand can easily be found.
  - Available clustering algorithm : KMeans
- 2. Choose algorithm which can be able to assign this previous label (group of electricity demand).
  - Available multilabels classifiers : KNN, SVM
- 3. Perform regression for each group of electricity demand using public data such as seasonal temperatures.
  - Available regressors : simple linear and Ridge using regularization and polyfeatures
- 4. Build a whole model (multiclassification and regression) for new raw data to predict future electricity demand and thus forecast electricity production to supply. General model should be similar to :

$$Total\ forecasted\ electricity\ demand = \sum_{i=1}^{Nb\ clusters} N_{users\ cluster\ i} * Predicted\ electricity\ demand_{regression\ cluster\ i}\ (feastures\ easily\ accessible)$$

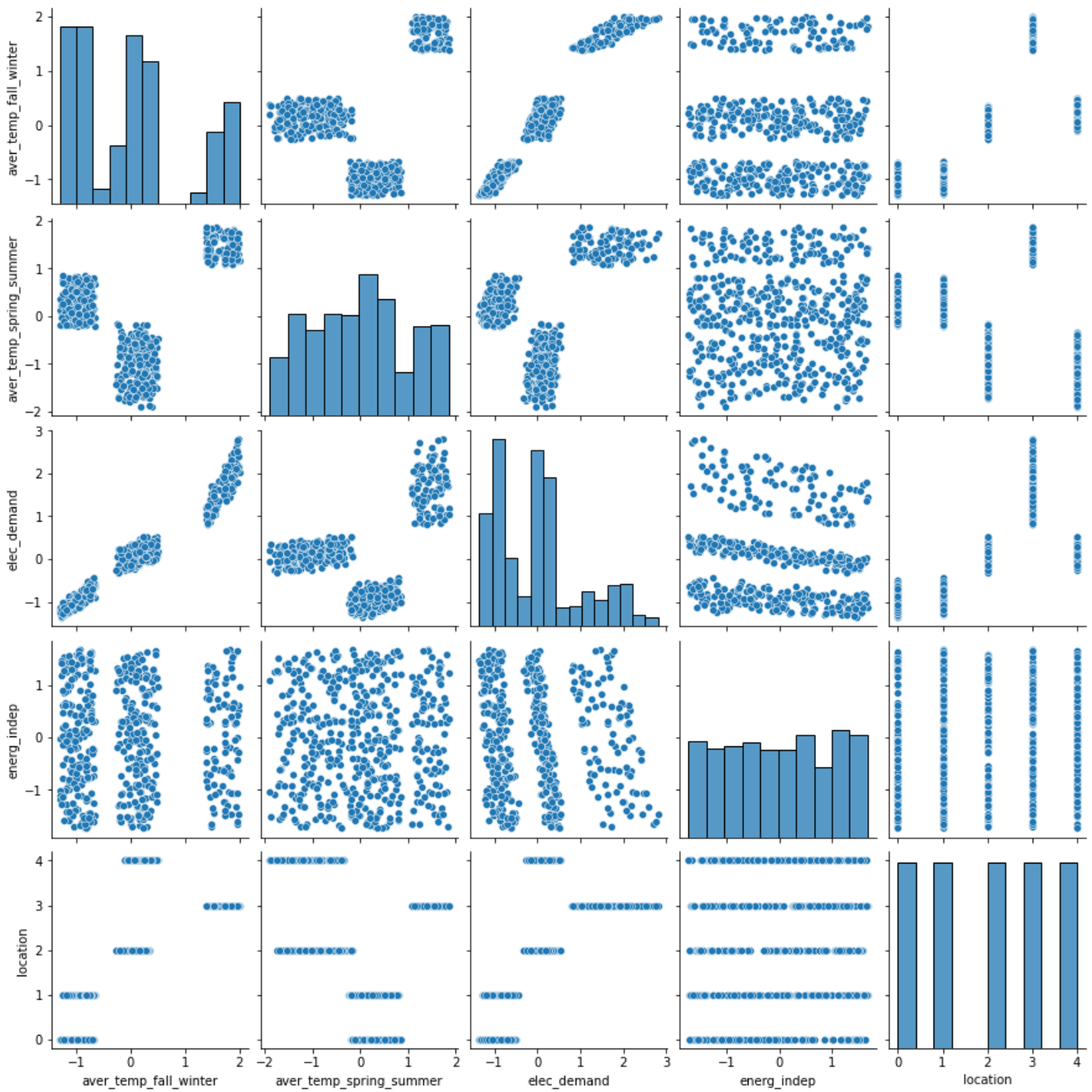
Our goal is to identify clusters, count users inside and build best regressors based on easily accessible features.

BEST PRACTICES

- Data are controlled to ensure non infinite value, no outlier, no multicolinearity between features.
- Algorithms are optimized on a training set (sub part of 80% of whole dataset after train_test_split method), assessed on a cross-validation set (sub part of 80% of whole dataset after cross_validate method) and finally applied on a final test set (20% of whole dataset after train_test_split method).
- Data are shuffled to build these train/CV/test sets in order no to have bias on results.
- Validation curves are plotted to see hyperparameters influence on performance.
- Learning curves are plotted to see training set size influence on performance.
- Algorithms are optimized to ensure a trade-off between bias and variance
  - Gridsearch are performed when less than 3 hyperparameters to explore
  - RandomSearch furthermore

CHECKING DATA

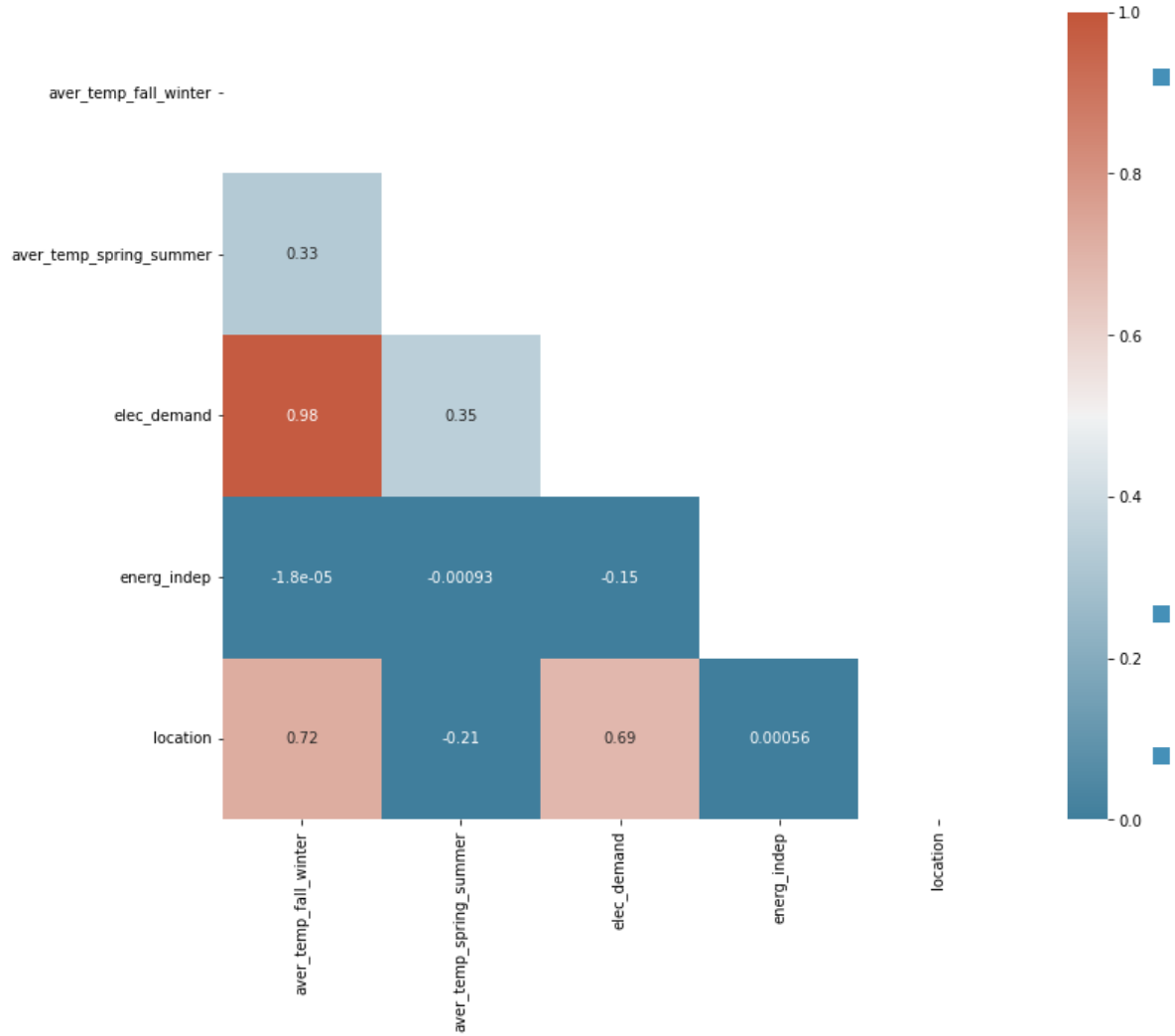
- Initial dataset : cleaned from outliers and missing data
- Location data (strings) : encoded
- Float data : normalized by StandardScaler() to ensure nominal computations for gradient computations



ANALYZING DATA

- ‘elec_demand’ is considered as a target variable. Final model will predict this parameter.
- Following as features : ‘aver_temp_fall_winter’, ‘aver_temp_spring_summer’, ‘energ_indep’, ‘location’
- Other sources of energy (energ_indep) : quite uniform distribution. Not very useful and not available from public data.
- Average temperature in spring-summer (aver_temp_spring_summer) : public parameter of interest to select.
- Average temperature in fall-winter (aver_temp_fall_winter) : public parameter of interest to select.
- Consumers area (location) : 5 distinct classes but not available in the future as a public parameter.

ANALYZING DATA

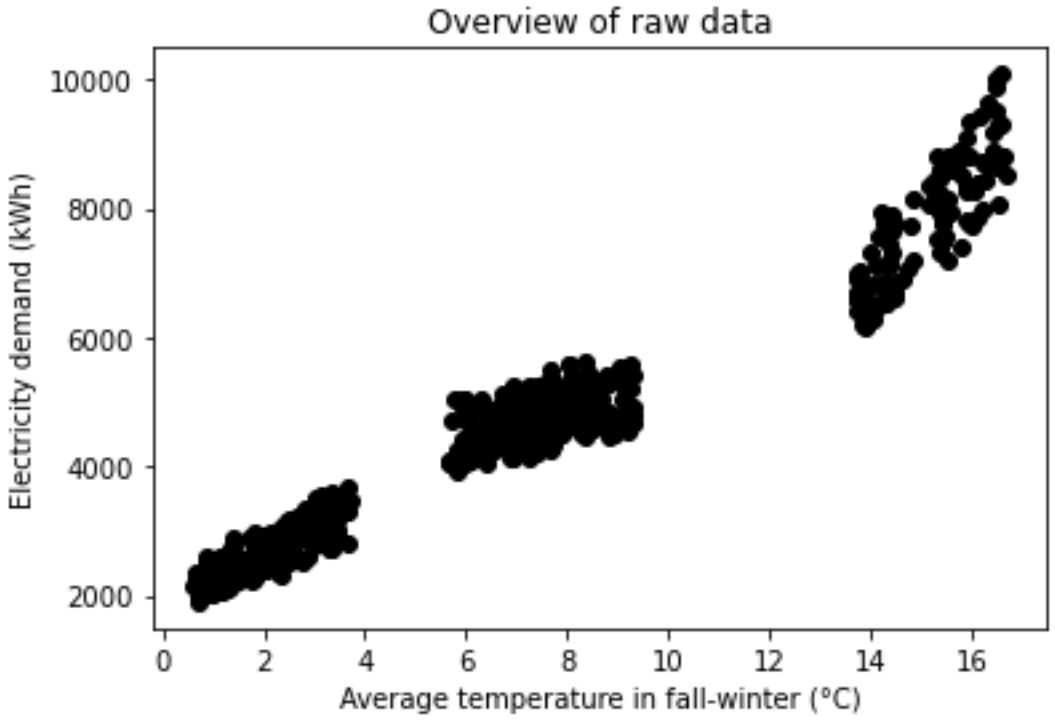


- No multicolinearity detected (VIF score < 10) : features are considered independent from each others (correlation matrix coefs < 0.75) and can be used to build the model.

Feature	VIF score
aver_temp_fall_winter	1.47507
aver_temp_spring_summer	1.23728
energ_indep	1
location	1.33464

- Energy independency will not be very convenient as it hard to get from future consumers.
- Features of interest for the final model are average temperatures in fall-winter (1st order influence on elec_demand) and spring-summer (2nd order parameter less correlated to elec_demand).

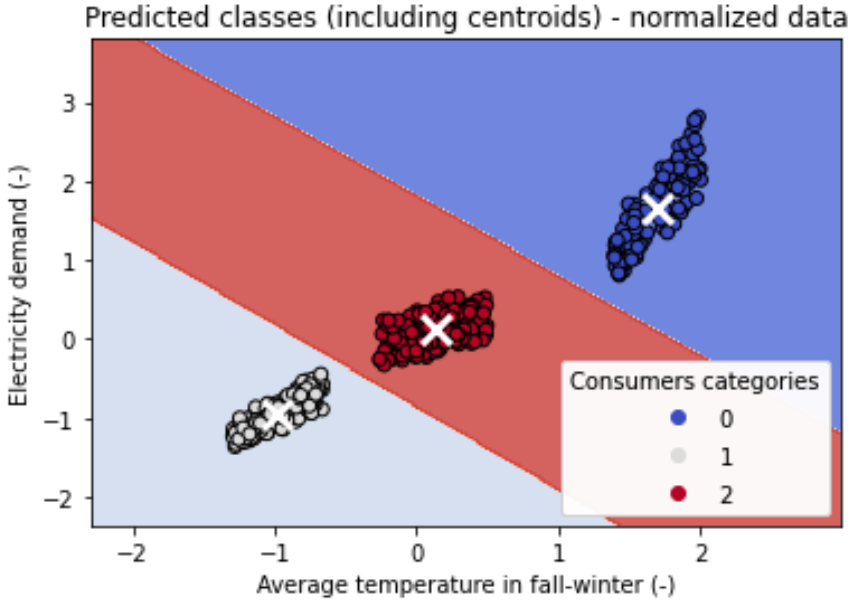
ANALYZING DATA



- Target variable is the electricity demand from consumers.
- Displaying this target after average temperatures in fall-winter is a good way to separate correctly three main groups of users : a low consuming class between 2000 and 4000 kWh, a medium one around 5000 kWh and a high last over 6000 kWh.
- This last feature will be considered in the next step to cluster data in those three groups.

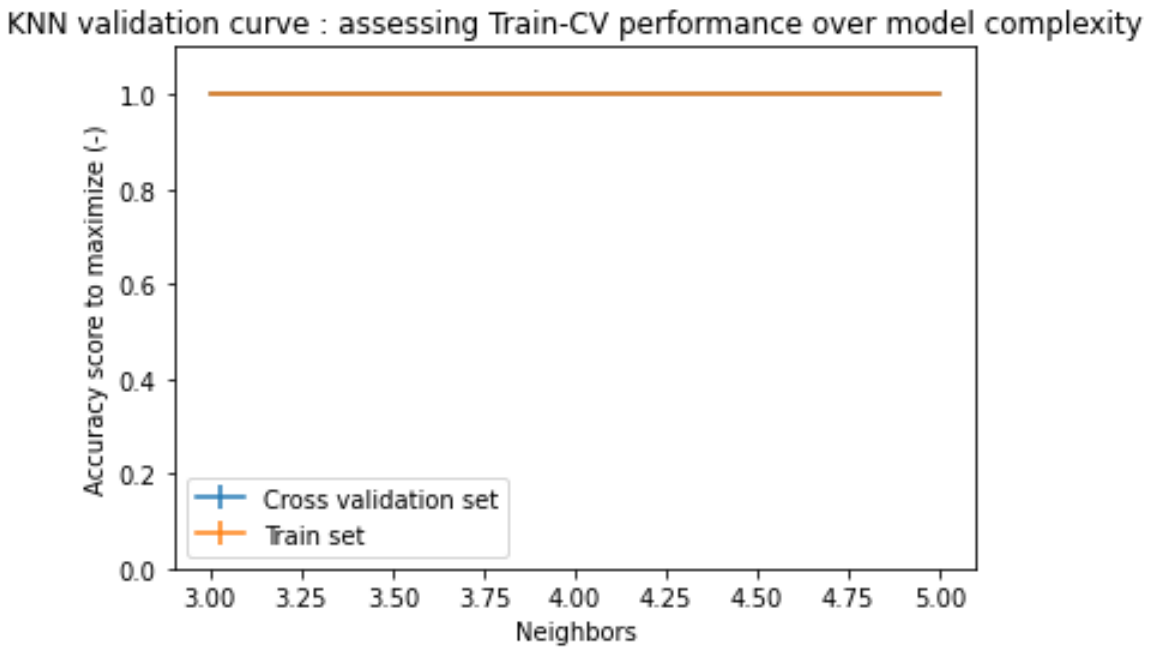
CLUSTERING

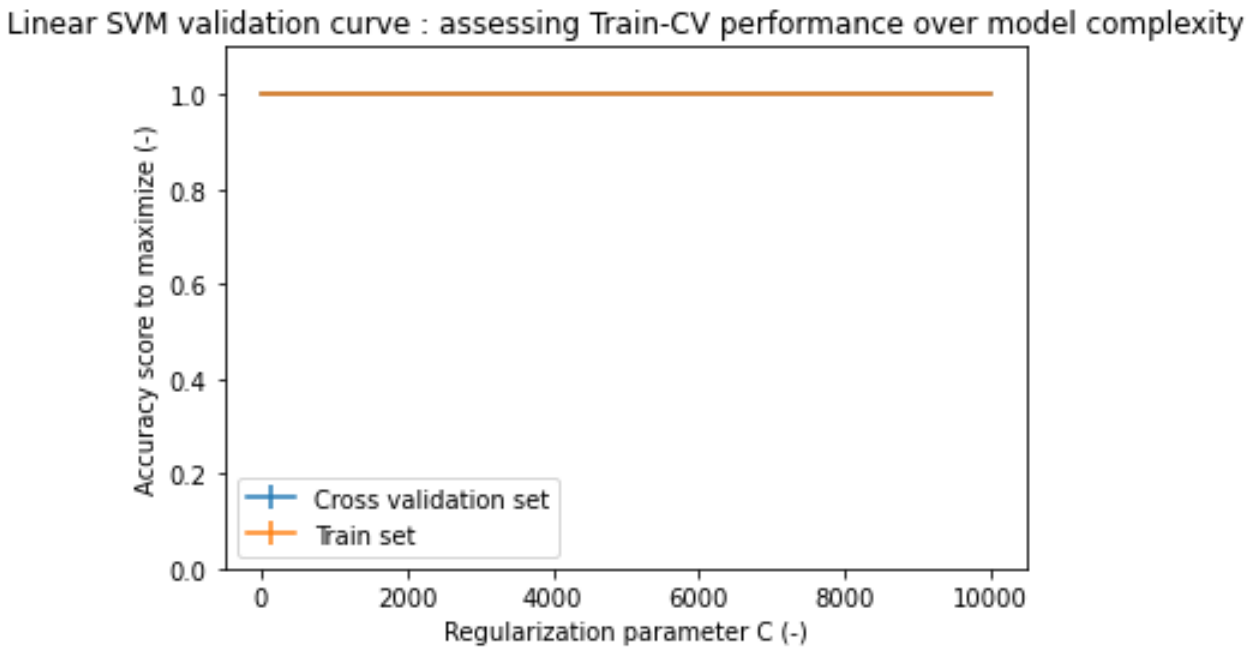
- KMeans algorithm computes groups of consumers (eg. labels, classes) after these variables :
  - Target : electricity demand
  - Feature : average temperatures in fall-winter
- A grid search finds the best number of clusters that maximizes silhouette criterion : it is found to be three as expected. They are labeled as consumers categories #0, #1 and #2.



MULTI-CLASSIFICATION

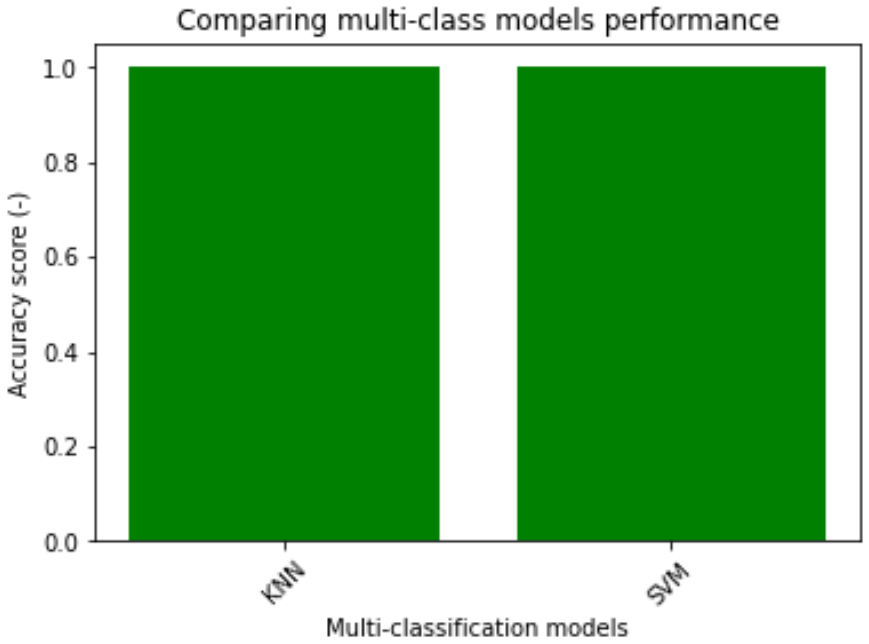
- Several algorithms perform grid search or randomized search to find optimal model that maximize accuracy score to predict consumers category for future unlabeled data.
- After future features (average temperatures in spring-summer and fall-winter), the multiclassifier will be able to predict consumers category (eg. Labels #0, #1 or #2).

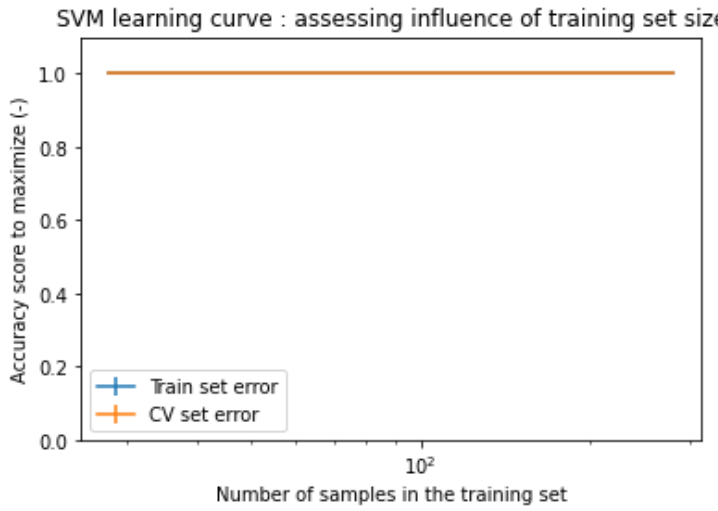


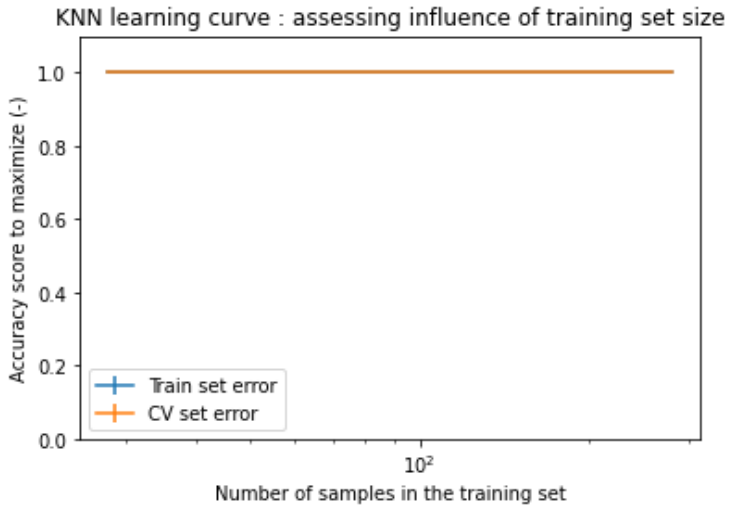


MULTI-CLASSIFICATION

- Such a perfect accuracy score on training data could be misinterpreted as an overfitting behaviour of algos : it is not the case as multiclassifiers are trained after well separated dataset, it is easy to get a very accurate score on both train and cross-validation sets.



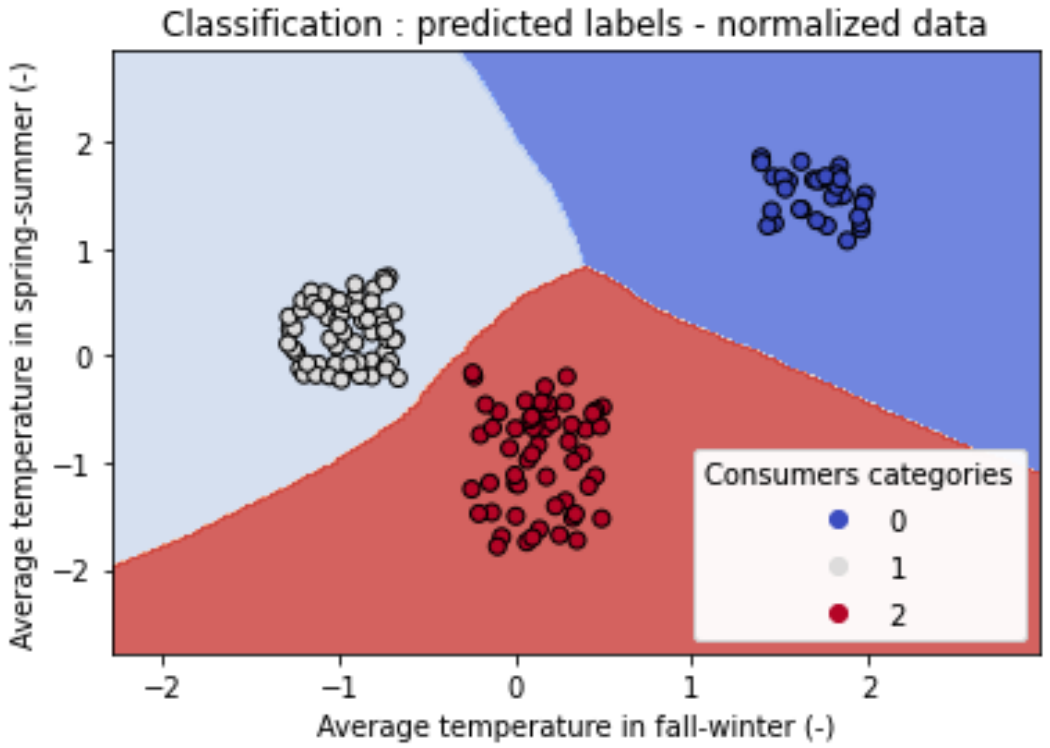




MULTI-CLASSIFICATION

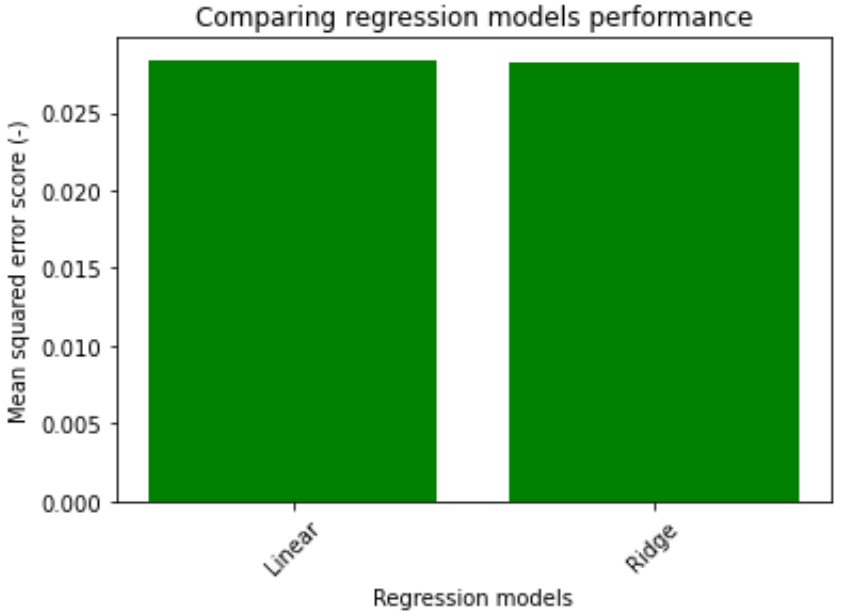
- KNN is chosen as multiclassifier algorithm with an accuracy score of 100% and an optimal number of neighbors of three as expected by clustering algorithm.

```
Classification : report on the test set:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        32
           1       1.00      1.00      1.00        62
           2       1.00      1.00      1.00        56

    accuracy                           1.00       150
```



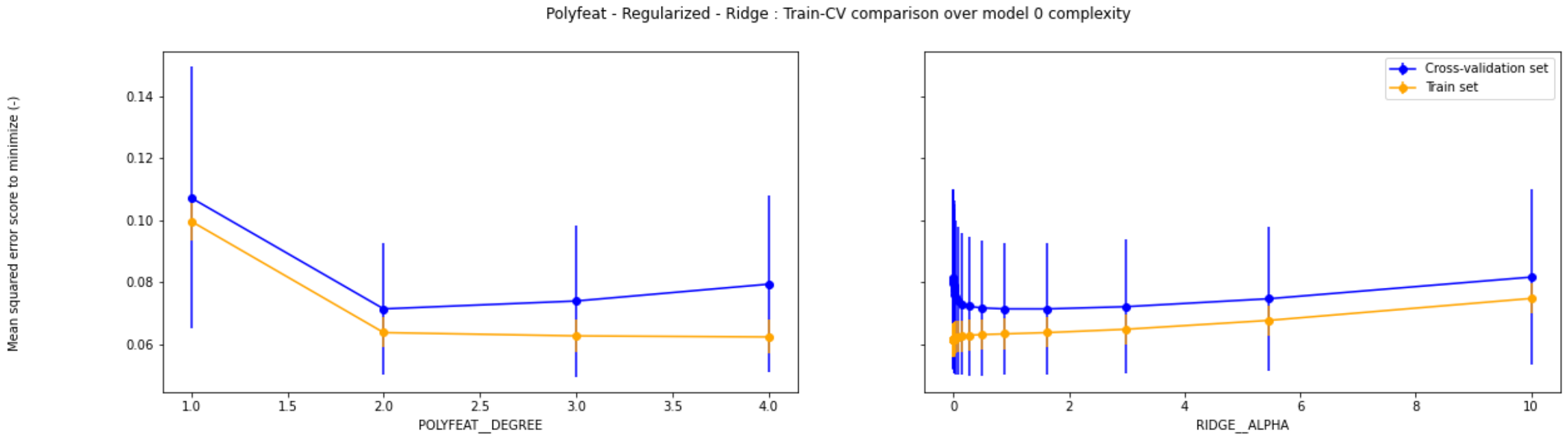
REGRESSION

- For each group of consumers (eg. Labels #0, #1 and #2), a regression model should be able to predict electricity demand (final target) after two features (average temperatures in spring-summer and fall-winter).
- Two regressors are assessed after mean squared error score to minimize (MSE = $RMSE^2$) : a simple linear aglorithm and a more complex one, a Ridge regressor using polyfeatures and regularization.
- A grid search finds optimal parameters for the ridge model : degree of polyfeatures (including degree=1 which corresponds to a linear case), and regularization parameter value (including low value as alpha=10^-4 which corresponds to a unregularized case). It is found to be the more accurate model for regressors #0, #1 and #2.



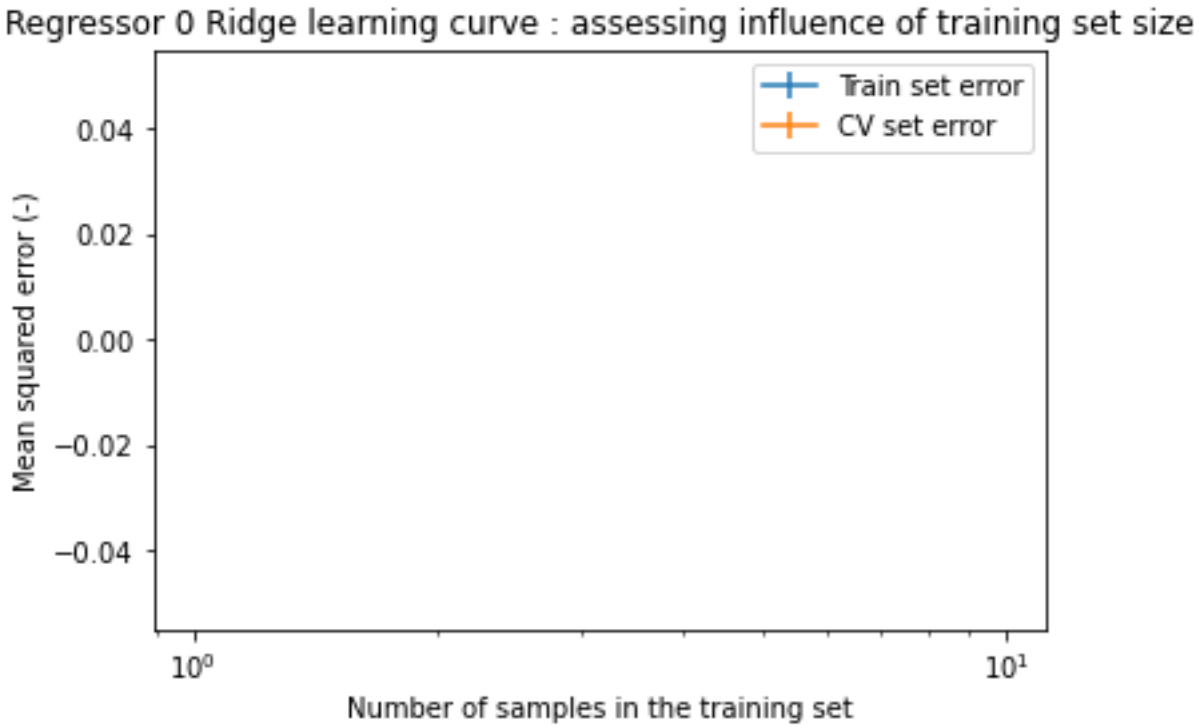
REGRESSION

- For example, here are validation curves for regressor #0 : both errors are low even if CV error is slightly bigger than train one. CV error is minimized for degree=2 of polyfeatures and alpha=1.6 as regularization parameter.



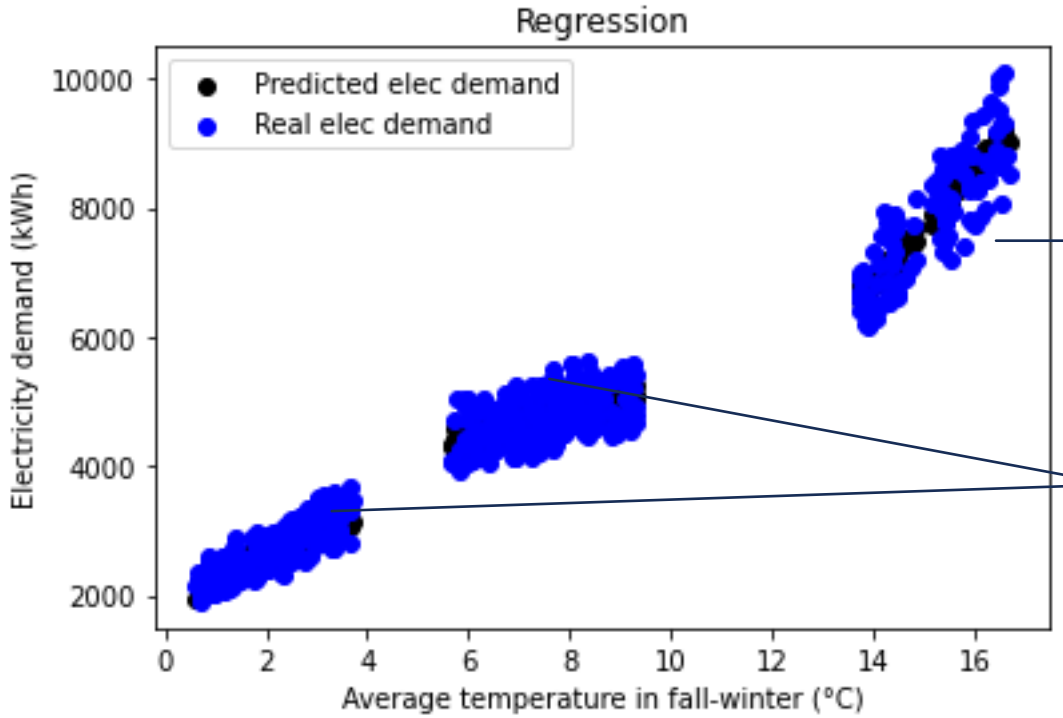
REGRESSION

- For example, here are learning curve for regressor #0 :



REGRESSION

- Each regressor concerns a group of electricity consumers (eg. Regressor 0 for label #0 and so on…). These three regressors are visible as three distinct stripes on the graph below.
- Predicted data fit very well test data.



This cluster seems to have a quadratic polynomial : confirmed by regressor with polyfeat=2

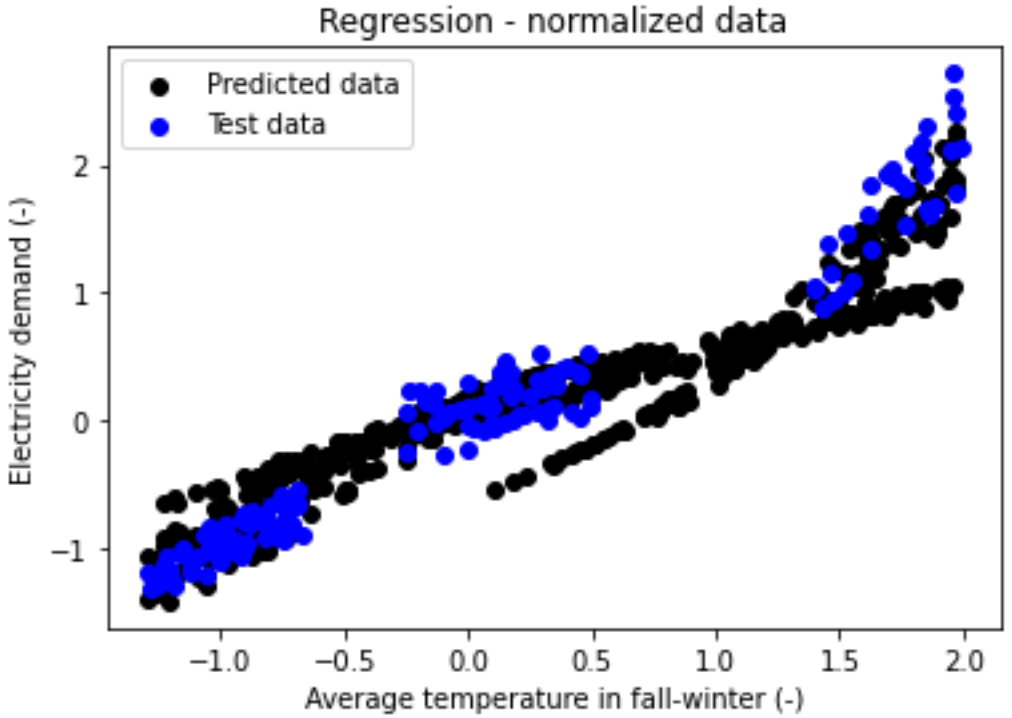
These clusters seem to have a linear polynomial : confirmed by regressors with polyfeat=1

FINAL MODEL

- Clustering model : KMeans algorithm identified very accurately (silhouette score of 0.75/1) three clusters based on whole dataset. A low consumption group (2000-4000 kWh/year) of users corresponding to East and Centre area (label 1), a medium one (around 5000 kWh/year) corresponding to North and West area (label 2) and a high consumption group (above 6000 kWh/year) corresponding to South area (label 0).
- Multiclassification model : KNN algorithm identfied obviously three neighbors as found by clustering model. It is able to find classes for future unlabeled data.
- Regression models :
  - Label 0 : Ridge using regularization (alpha=1.6) and polyfeatures (degree=2)
  - Label 1 : Ridge without regularization (alpha=0.0) nor polyfeatures (degree=1)
  - Label 2 : Ridge using low regularization (alpha=0.3) and no polyfeature (degree=1)
- New data :
  - Features available : average temperatures in fall-winter and spring-summer
  - Model able to predict electricity demand as a target

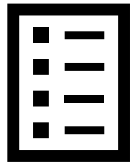
PREDICTING FUTURE ELECTRICITY DEMAND

- Test data come from initial dataset
- Predicted data are built after test data as plain intervals of [min(aver_temp_fall_winter);max(aver_temp_fall_winter)] and [min(aver_temp_spring_summer);max(aver_temp_spring_summer)]
- Three black stripes appear outside blue groups as they have been identified by the final model as belonging wether to label 0 or 1 or 2. They can been seen as extrapolation of blue groups as they are wider.



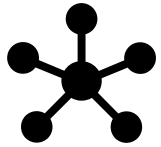
CONCLUSION

Initial dataset representative of general trends for electricity consumption

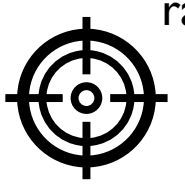


Dataset analyzed and features of interest identified

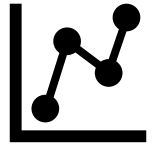
Three groups of consumers found by clustering



Multiclassifier learnt to predict groups of consumers for future raw data (unlabeled)



Three regressors predict electricity consumption after basic features (average temperature in spring-summer and fall-winter) found in future raw data



Basic data for new consumers (average temperatures) are sufficient to predict future electricity consumption

An better electricity production based on…

…an accurate forecasted consumers demand