Acta Polytechnica CTU Proceedings **38**:90–96, 2022
https://doi.org/10.14311/APP.2022.38.0090

Published by the Czech Technical University in Prague

ENERGY PERFORMANCE ESTIMATION FOR LARGE BUILDING PORTFOLIOS WITH MACHINE LEARNING-BASED TECHNIQUES

Frédéric Montet[a,*], Alessandro Pongelli[b], Jonathan Rial[a], Stefanie Schwab[c], Jean Hennebert[a], Thomas Jusselme[b]

[a] *University of Applied Science of Western Switzerland (HEIA-FR, HES-SO), iCoSys Institute, Bd. de Pérolles 80, 1705 Fribourg, Switzerland*

[b] *University of Applied Science of Western Switzerland (HEIA-FR, HES-SO), ENERGY Institute, Bd. de Pérolles 80, 1705 Fribourg, Switzerland*

[c] *University of Applied Science of Western Switzerland (HEIA-FR, HES-SO), TRANSFORM Institute, Bd. de Pérolles 80, 1705 Fribourg, Switzerland*

[*] corresponding author: frederic.montet@hefr.ch

Abstract. Building operation is responsible for 28 % of the world's carbon emissions. In this context, establishing priorities in refurbishment strategies at the scale of a city or a group of buildings is important. Such procedures are usually led by experts in energy performance and, therefore, they are rarely carried out due to their long and costly nature.

This research aims at the estimation of building energy performance to pave the way towards finding near-optimal refurbishment strategies. Thanks to the identification of easily-accessible building characteristics, the method applies machine learning models to scan a building portfolio based on a low level of details. The results show good potential to identify low-performer buildings with simple machine learning methods. It also opens the door for further improvements through the inclusion of supplementary building features at the input of the predictive system.

This work includes (a) the integration of a knowledge database thanks to the Swiss CECB energy performance certificates, referencing more than 70 000 buildings, (b) the preparation of a training data set through the selection of relevant physical characteristics of buildings (input) and the corresponding energy consumption labels (output), (c) the development of predictive models used in a supervised way, (d) their evaluation on an independent test set.

Keywords: Refurbishment strategies, machine learning, energy performance certificates.

1. Introduction

In a world where climate change imposes major adjustments in order to slow down the rise in environmental temperature, one important area to improve is the planning of building renovations. Thanks to changing construction techniques and equipment, building renovations became a major focus of energy strategies of governments.

Switzerland, for example, is implementing a strategy called "Energy Strategy 2050" which aims, among other things, to reduce the energy consumption of buildings through incentives [1]. This, in the coming years, may cause a rush to renovate old buildings and to prioritise the management of renovations due to the costs and timeframes imposed. For building stock managers it means having to priorities and create a ranking of interventions that can lead to energy improvements, but which do not involve high costs. In Switzerland there is the Cantonal Energy Certificate for Buildings (CECB) [2], which allows the assessment of the current state of the building and the planning of a possible renovation. It allows to attribute an energy label to the efficiency of the building envelope, which describes the quality of the thermal envelope including roof, wall, floor and window insulation, and also takes into account thermal bridges and the shape of the building. A second label is given to the overall energy efficiency including heat demand, electricity demand, own production of electricity as well as the building's equipment for heat and domestic hot water. The labels are divided into 7 classes: from A, the best class, to G, the worst class compared to a reference. This certificate is compulsory in cantons such as Geneva, Vaud, Fribourg, Neuchâtel, Nidwald, Zug and Zürich, in case of sale and/or renovation of the building. This has led to the creation of a database with more than 70 000 certificates describing the physical characteristics and energy performance of each building.

However, this certificate oblige experts to travel to the building site in order to collect the various data needed for in-depth analysis, increasing the cost and time required to analyse a large building stock. In recent years it has also become possible, through portals such as Registre fédéral des bâtiments et des logements (RegBL) [3], Cantonal geoportals [4, 5], Google street view [6] and Google maps [7], to find precise data or to collect it remotely. Therefore, in the

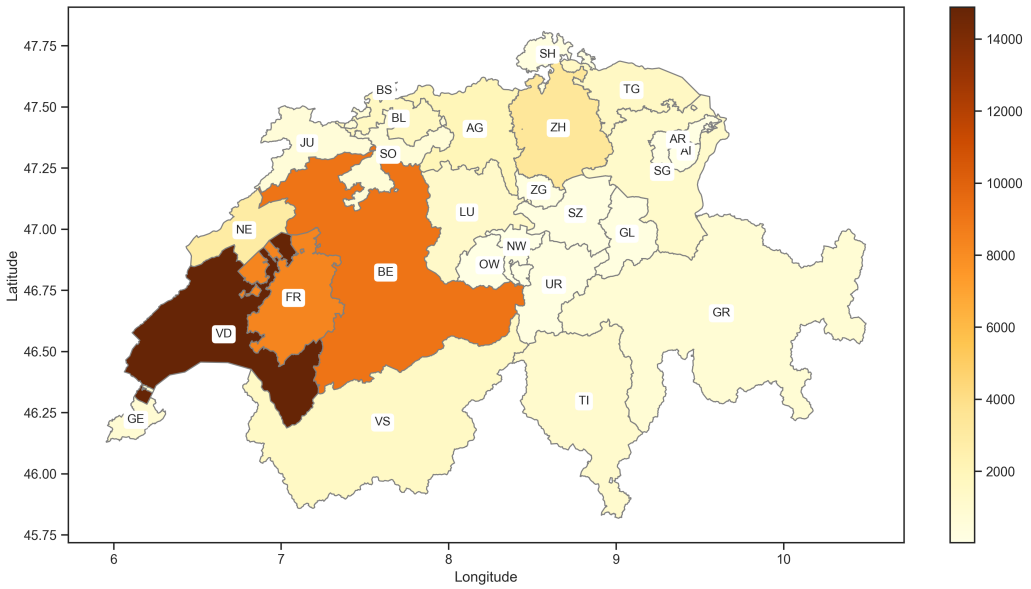


FIGURE 1. Map of Switzerland showing the concentration of building energy performance certificates (CECB) per cantons.

following chapters we will try to answer the main question: "How can the available online data be used to quickly classify a building stock energy performance?".

In this paper, we investigate the use of Machine Learning (ML) approaches as a solution to estimate automatically the less performing buildings according to the CECB methodology. The proposed method is to take as input of the ML systems easy-to-find building characteristics that do not need the intervention of an energy expert. The output of the system is a prediction of the CECB energy label of the building. If functional, such approaches could provide a quick and easy way to rank buildings by priorities of renovations. The CECB association gave us access to his data, under a data agreement for the protection and sharing of sensitive data. We used these data to train and test our ML systems.

The paper is organized as follows. Section 2 presents the methods used to prepare the data and to select and optimize the best performing models. Section 3 presents the obtained results according to specific metrics that we propose to evaluate their performances. Finally, Sections 4 and 5 present discussions, conclusions and future works.

2. METHODS

This section introduces exploratory and preprocessing phases performed on the data. Details regarding the machine learning models selection, their optimization and the process used to assess the results quality are provided.

2.1. DATA EXPLORATION

As indicated in the previous Section 1, more than 70 000 certificates are in the dataset. To understand the nature of the data, statistics were computed as a preliminary analysis. Knowing that in some cantons the certificate is compulsory, it was verified that this is reflected in the form of certificates in the database. In Figure 1 it is possible to see how the dataset is divided according to the different cantons and it is possible to identify the cantons that have a compulsory certificate.

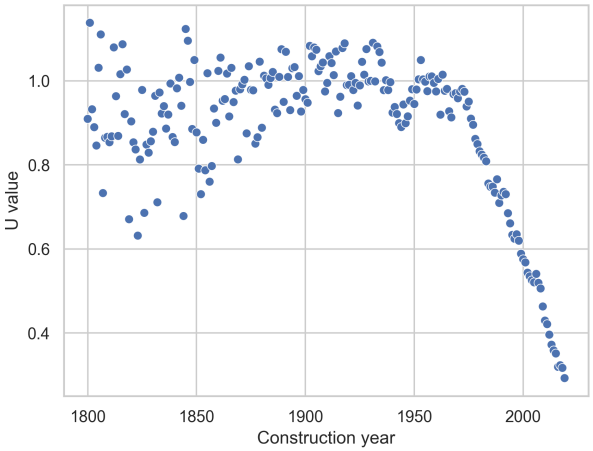


FIGURE 2. Evolution of the average heat transfer coefficient in relation to the year of construction of buildings.

A second analysis was carried out to verify data consistency. In Figure 2, it is possible to see the average heat transfer coefficient for each year. The evolution of this value is getting lower through time, which correlates with the increase in building insulation from 1970. This conclusion is similar to the findings of the Energy and Renovation (eREN) project [8].

To get a better understanding of the subdivision of the dataset, we then checked which categories are present. They are subdivided according to the cat-

Categories	CECB Qty.	Percentage
Single-family building	40 590	55.10 %
Multi-family building	27 780	37.71 %
Administration	1 932	2.62 %
Mixed	1 802	2.45 %
School	1 374	1.87 %
Retail	79	0.11 %
Hotel	77	0.20 %
Restaurant	29	0.04 %

Table 1. Representation of the various building categories in the CECB dataset.

	Original data	After cleaning
A	1620	277
B	10695	4594
C	16036	10056
D	17271	10718
E	11615	7355
F	6950	4374
G	9474	4986

Table 2. Number of certificates per label before and after cleaning the dataset.

egories of RegBL [3]. In the Table 1, it is possible to see that the two most represented categories are single-family building and multi-family building.

The subdivision in classes before the cleaning of the dataset is present in the Table 2 where we can see that the most represented class is D followed by other classes and ended by the least represented A.

2.2. Data Preprocessing

The data received was in the form of several tables. To use them, the latter were merged to combine the data in a single structure. Subsequently, a data cleaning was carried out. After manual checks, outliers were identified and removed when samples were outside two standard deviations from the mean, thus removing 5 % of the data. In addition, missing values were removed when their percentages were more than 90 % per column or more than 80 % per samples, i.e. per certificate.

The result of this part of the method is presented in Table 2 as a number of certificates after the cleaning.

For all following machine learning (ML) tasks, a balanced dataset is preferred[1]. To achieve this, sub-sampling can be used. Since the class A is clearly underrepresented with its 277 samples, sub-sampling the data would induce a high loss of samples. For this reason, class A and B are merged before sub-sampling for all further ML tasks. Moreover, as A and B building classes are best performers, they are not targeted by refurbishment plans.

[1] A dataset is said balanced when the number of samples per class is equal.

2.3. Model Performance Exploration

The modelling method includes 3 steps.

(1.) Pre-processing routines were carried out. The latter include an ordinal encoding of the data as well as a normalization by computing the Z-score on all variables.

(2.) Then, eighteen different models from the Scikit-learn library were selected when appropriate for a classification task and trained with a k-fold cross validation with $k = 3$ [9].

(3.) The evaluation of all models is made based on their F1-scores and accuracy metrics to select the two most promising ones.

2.4. Optimization and Model Selection

On the two best models from the previous section, optimization and training is performed to allow for final model selection. Since the optimization tasks is memory expensive, the research for optimal parameters has been done on a reduced number of samples per class selected in a random manner.

The optimization includes a cross-validated randomized search where $k = 3$ on a selection of parameters (see Section 3). Model training is computed as in step 3 from Section 2.3. Finally, the overall best model is selected given its per-class metrics (Accuracy and F1) and confusion matrix.

2.5. Special Cases Identification

As a last step of the method, a more in-depth analysis of special cases is performed to identify the reasons behind a high prediction error.

A distance between classes is computed with a method taken from ordinal regression problems [10]. To compute this distance, let $y = A$ and $\hat{y} = G$ be a sample's class and its estimation. Their encoded version would be $y' = (1, 0, 0, 0, 0, 0, 0)$ and $\hat{y}' = (1, 1, 1, 1, 1, 1, 1)$. The distance d between the two classes can then be calculated with $d = \sum_{i=1}^{n} |y'_i - \hat{y}'_i|$.

Once obtained, all samples were ordered by descending order from their distance d.

3. Results

3.1. Model Performance Exploration

The eighteen algorithms were trained on a resampled dataset to make it balanced. Performances obtained in the Table 3 allow for the comparison of all algorithms in order to identify the two best candidates for energy performance certificates predictions.

The `DummyClassifier` performs a classification in a random manner with 0.17 accuracy. The latter result sets a baseline above which an algorithm learns the characteristics of the data. Since most algorithms performed with a F1-score and accuracy between 0.3 and 0.5, the problem at hand can be characterized as difficult.

Model	Accuracy	F1-score
HistGradientBoostingClassifier	0.50	0.49
RandomForestClassifier	0.50	0.48
ExtraTreesClassifier	0.49	0.48
BaggingClassifier	0.49	0.48
GradientBoostingClassifier	0.49	0.47
MLPClassifier	0.47	0.46
SVC	0.46	0.45
NuSVC	0.46	0.45
LogisticRegression	0.45	0.43
DecisionTreeClassifier	0.43	0.43
AdaBoostClassifier	0.43	0.42
LinearDiscriminantAnalysis	0.43	0.41
ExtraTreeClassifier	0.39	0.38
KNeighborsClassifier	0.39	0.38
LinearSVC	0.41	0.37
SGDClassifier	0.40	0.35
RidgeClassifier	0.39	0.34
NearestCentroid	0.35	0.34
Perceptron	0.33	0.32
PassiveAggressiveClassifier	0.32	0.31
BernoulliNB	0.32	0.31
GaussianNB	0.30	0.30
DummyClassifier	0.17	0.05

TABLE 3. Model performances exploration. Models are ordered by F1-scores and then, by their accuracy metrics.

	Before	After	Gain
Catboost	0.59	-	-
HistGradientBoosting	0.49	0.58	18 %
RandomForest	0.48	0.57	19 %

TABLE 4. Comparison of the best classifiers with their F1-score average across class.

The top 5 algorithms include recent classification models performing with similar result around 0.48 F1-score. The `MLPClassifier`[2] performed with moderately good result. The best selected candidates for the next step of the method are the `HistGradientBoostingClassifier` and `RandomForestClassifier`.

3.2. Optimization and Model Selection

The fine tuning procedure was performed on `HistGradientBoostingClassifier` and `RandomForestClassifier` models. In addition, one state-of-the-art gradient boosted tree implementation was added from the Catboost library, without fine-tuning [11].

As Table 4 shows, all models are reaching performances above or close to 0.5 without optimization, thus already allowing for an educated guess. When parameters from Listing 1 are used, substantial gains can be obtain within the order of ∼20 %; thus proving the value of randomized search.

[2]Stands for Multi-layer Perceptron, i.e. a neural network with the default parameters from the Scikit-learn library.

```
best_params_hist_gradient_boosting = {
    'l2_regularization': 0.00081,
    'learning_rate': 0.11889,
    'max_bins': 60,
    'max_leaf_nodes': 11,
    'min_samples_leaf': 87
}

best_params_random_forest = {
    'n_estimators': 800,
    'min_samples_split': 5,
    'min_samples_leaf': 1,
    'max_features': 'sqrt',
    'max_depth': 80,
    'bootstrap': False
}
```

LISTING 1. Best parameters.

	Unbalanced	Balanced
AB	0.76	0.82
C	0.69	0.62
D	0.60	0.54
E	0.45	0.44
F	0.31	0.47
G	0.68	0.67
Accuracy	0.60	0.59
F1 avg.	0.59	0.59

TABLE 5. Global building efficiency F1-scores.

A final F1-score average across classes of 0.59 makes the Catboost based model the more performant of the selection. Also, both algorithms from Scikit-learn library had a notably high performance gain, making them reach a scores comparable to Catboost.

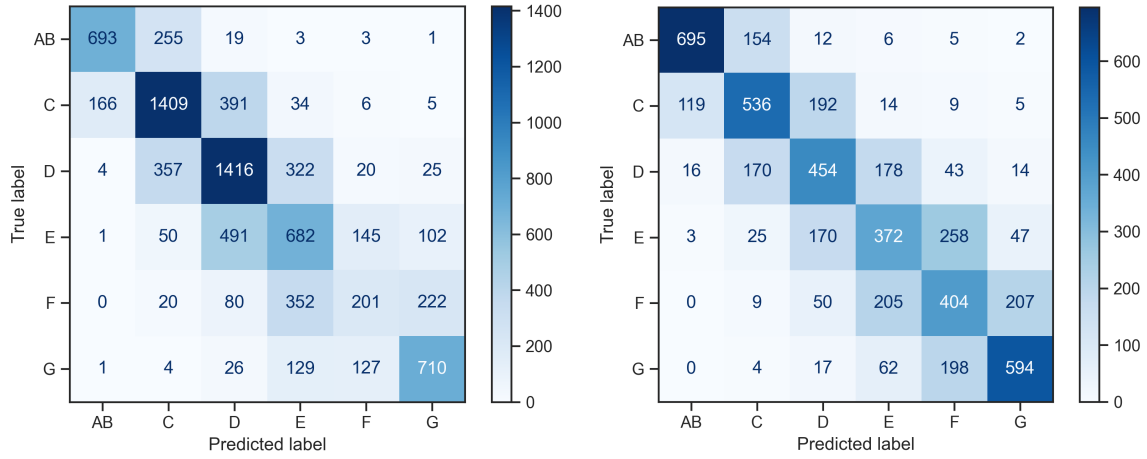
3.3. Best Model Performances

The `Catboost` model, using gradient boosting on decision trees, is the final model selected. This section introduces it in three phases. First, with explaining an additional data processing task. Second, by presenting the model performances in depth. Finally, by analyzing the errors produced by the model.

3.3.1. Model Performance

The final model performances are summarized in Table 5. In average, the model reaches an accuracy of ∼0.6. This represents an improvements of ∼350 % compared to the baseline of 0.17 given by the `DummyClassifier` from Section 3.

On unbalanced data, scores have a great variability, which makes the model quality hard to evaluate with confidence. Since unbalanced data has generally more samples, the awaited behavior is a higher score, but this isn't the case for all classes. A possible explanation lies in the selection of more representatives samples while sub-sampling.



(A). Unbalanced dataset. (B). Balanced dataset.

Figure 3. Confusion matrices for global efficiency.

On balanced data, the comparison between classes gives an insight on the difficulty to evaluate each class. From AB to F, scores are decreasing from 0.82 to 0.47. This shows that the lower the efficiency of a building is, the harder it becomes to predict its class. Both end of the class spectrum – AB and G – have higher scores. This behavior is probably due to the absence of strict intervals i.e. a certificate made on a highly inefficient building would have a class higher than G, but since it doesn't exist, it is classed as G.

More details about the predictions uncertainty are given by the confusion matrices in Figure 3 where the precision of the models are presented visually. A highlighted diagonal is predominant in both the unbalanced and balanced plots; showing the correct predictions. On each side of the diagonal, the predictions made with an error of one class above or below the true label are represented.

3.3.2. Prediction Reliability

To evaluate how the false predictions are spread around the diagonal, the plot on Figure 4 shows the cumulative density of the distances between predicted and true value. The accuracy of ∼60 % is visible where the distance is 0. Then, in case of a wrong prediction, there is a ∼90 % probability that the true class is only a letter away; ∼100 % for two letters, and so on. This highlights that wrong predictions are generally not far away from their target.

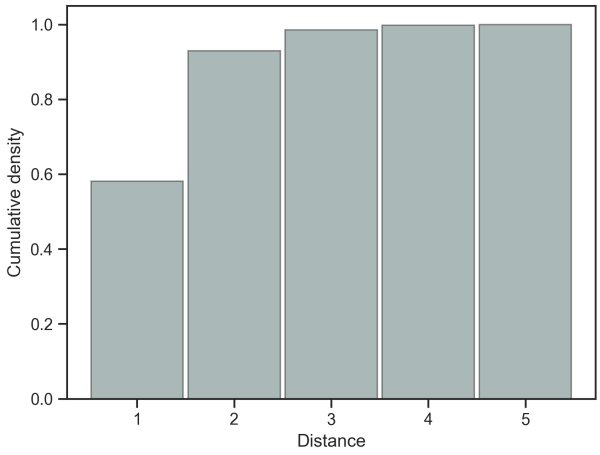


Figure 4. Cumulative distribution of the distances between true and predicted global efficiency given two two different metrics.

3.4. Online variables

To assess whether the necessary input variables are available online to speed up the classification process with ML-based techniques, a wide search was carried out on the various portals listed in the Section 1. The Figure 5 presents the most discriminant variables in the Catboost algorithm. For brevity, only the ten firsts are presented.

Of the most important variables for the operation of the algorithm, we can easily find the year or era of construction of the building on the RegBL site through the interactive building map. A second variable that can easily be found is the building width through Google Maps for example as the measurement tool can be used to obtain a value of the desired building.

A third and a fourth that can be found on the internet, but is subject to restrictions, are the year of construction of the energy agent and the energy source. They are available in the RegBL database, but this type of data is only granted after authorization and verification of credentials.

All other variables are not accessible as they are not present on any platform at the moment.

4. Discussion

The applicability of the methods depends on

(1.) the accuracy of the model that is used to predict performance classes, and

(2.) the easy online access of inputs of the predicting model describing the physical properties of buildings to be classified.

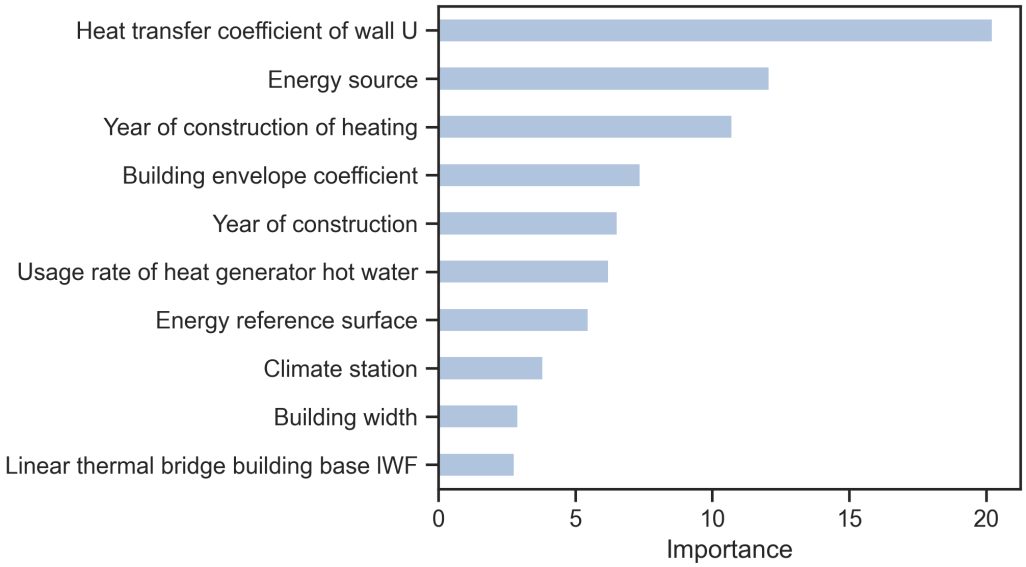


FIGURE 5. Feature importance of the first ten variables used by Catboost.

Regarding (1.), the Catboost model is promising as it has close to 60 % accuracy. Moreover, the prediction is only 1 class away in 90 % of the buildings. This seems to be highly acceptable if the method decreases dramatically the time spent to assess the building performance.

In order to be able to understand how to further improve prediction, it is necessary to start by analysing the process used step by step, starting with the collection of data for the generation of the CECB certificate.

The initial dataset is the result of certificates executed by many CECB experts. Several points in the creation of a CECB certificate are decided through visual inspection and thus based on the knowledge of the individual expert. For example, in order to get a U-value for walls, the expert can enter a proposal for the composition of the wall by visual inspection and the program calculates a U-value based on what the expert insert. This means that each expert, according to his knowledge, can enter or omit data, but in the end the program still manages to generate a complete certificate. This, for example, explains the missing data for some certificates. In order to have a more complete dataset, it would be important to reduce the possibility of omitting data for the generation of a certificate to a minimum.

Regarding the type of certificate, before 2012, a simpler type of certificate with fewer parameters was in place, then a more detailed certificate was adopted. This certainly causes a lack of data on some of the certificates used. Filling the gaps and convert less detailed certificates into more detailed ones by entering the missing data could be a solution to improve the base set.

During the data cleaning, arbitrary choices were made to eliminate values with the 3 sigma method, to eliminate columns with 90 % of the data missing as well as certificates with 80 % of the data missing. This system can clearly decrease the accuracy by removing variables and especially population from the data set. A more thorough analysis for cleaning the data should be done, so as to be sure that all inconsistent values are removed and that all consistent values are kept. The same can be done by keeping variables considered important as well as certificates to increase the population.

Finally, when it comes to prediction, changing the settings of the algorithm could lead to further improvement of the dataset. A better tuning can lead to a more precise prediction.

Moving on to the second point (2.), it was identified in Section 3.4 that only two variables are easily accessible. In the next part of the article we will analyse each variable in the Figure 5, exploring possible methods of providing representative values of the analysed building for the data required.

In order, the most important variable is the heat transfer coefficient of walls, which cannot be found online. One solution would be to send a person on site to do an analysis in order to give a value to this variable. A second possibility that should be explored is to use artificial intelligence to reconstruct this value, for example from an image of the wall of the building and the year of construction, so that it can recognize some key features and then give a value based on other similar buildings.

Moving on to the second variable in the list is the Energy source. As already discussed in the Section 3.4, it can be found on the RegBL website with restrictions. However, this data may be in the possession of the owners or managers of the building and could therefore be entered easily.

The year of construction of the heating, as mentioned in the Section 3.4, can be found on the RegBL website with restrictions. Also this variable should be checked if the owner or manager of the building is in

possession of it.

The envelope coefficient is not found directly but could be calculated from other variables. Online it is possible to measure the building perimeter through Google Maps. On the RegBL website it is possible to find the number of floors in the building and the ground area. Using for example an average value for the height of a floor, the various parameters can be combined. Multiplying the perimeter by the number of floors and the average height and then adding the floor area twice gives an approximation of the envelope surface. Dividing the envelope surface by the number of floors multiplied by the floor surface gives an approximate value of the envelope coefficient. Clearly an approximate value that should be checked for potential and especially for possible errors brought with it.

The utilisation rate of heat generator for hot water cannot be found online and so there are two possibilities to find this data. The first is that the manager or owner of the building knows this value. The second is to send an expert on site to make an assessment.

The energy reference surface can also be estimated from other variables. On the RegBL website we find the floor area of the building and the number of floors, multiplying these variables together we find a rough estimate. Clearly, the reliability of predicting the correct energy class using this approximation must be checked.

The climate station is not available online, but knowing the address of the building it is possible to indicate which weather station is to be used for the calculation.

The linear thermal bridge building base lWF is like the heat transfer coefficient of walls, is not available online and it is possible to use the same solution proposed.

It must be said that all missing variables could be estimated or found easily with an expert on site. Having only 10 variables to find would simplify and speed up the work to be done on site.

5. Conclusion

In this paper we have highlighted the preliminary reliability results of using a classification algorithm to analyse a building. The result of ranking the building in the good class with 60 % accuracy is a promising result for future developments. It should be noted that there is a 90 % probability of being in the adjacent class, which brings value to the work done.

In addition, the work carried out to check the online presence of the most important variables for prediction has shown that it is still premature to find the exact value online. Nevertheless, it may be possible to recreate some them by developing further techniques.

Some recommendations for future work are necessary. For example, A special attention must be taken when merging the different data available during the preparation phase, as this could lead to consecutive errors in the other phases of the work. A special attention must also be taken during the cleaning phase to ensure that the maximum amount of data is available to carry out the work.

Acknowledgements

The authors would like to express their gratitude to Ms Karine Wesselmann and the CECB association for providing the data for this work.

They would also like to thank Professor Mylène Devaux and the iTEC institute for their collaboration.

Financial support is gratefully acknowledged from the HEIA-FR Smart Living Lab research program.

References

[1] Swiss Federal Office of Energy. Stratégie énergétique 2050, 2020. [2021-11-05]. https://www.bfe.admin.ch/bfe/fr/home/politik/energiestrategie-2050.html

[2] Association GEAK-CECB-CECE. Le Certificat énergétique cantonal des bâtiments (CECB). [2021-11-05]. https://www.cecb.ch/

[3] Office fédéral de la statistique. Registre fédéral des bâtiments et des logements (RegBL). [2021-11-05]. https://www.housing-stat.ch/fr/index.html

[4] Canton de Fribourg. Portail cartographique du canton de Fribourg. [2021-11-05]. https://map.geo.fr.ch/

[5] Canton de Neuchâtel. Portail cartographique du canton de Neuchâtel. [2021-11-05]. https://sitn.ne.ch/

[6] Google LLC. Google Street View. [2021-11-05]. https://www.google.com/intl/en_ch/streetview/

[7] Google LLC. Google Maps. [2021-11-05]. https://www.google.ch/maps/

[8] S. Schwab, L. Rinquet, M. Devaux, et al. Rénovation énergétique. Tech. rep., Fribourg, Suisse, 2018.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**:2825–2830, 2011.

[10] J. Cheng, Z. Wang, G. Pollastri. A neural network approach to ordinal regression. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1279–1284. 2008. https://doi.org/10.1109/IJCNN.2008.4633963

[11] L. Prokhorenkova, G. Gusev, A. Vorobev, et al. CatBoost: unbiased boosting with categorical features, 2019. arXiv:1706.09516