Utrecht University

Missing Data in Real Estate: Comparing different machine learning algorithms to predict a real estate object's energy label and building type

Luitwin Mallmann

FACULTY OF SCIENCE

2021

Missing Data in Real Estate: Comparing different machine learning algorithms to predict a real estate object's energy label and building type

LUITWIN MALLMANN

FACULTY OF SCIENCE

**A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN APPLIED DATA SCIENCE**

2021

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

Table of Contents

Abstract

This master thesis aims to present the successful application of a variety of machine-learning algorithms and model building strategies to predict the building types of Dutch real estate objects available on ”kadaster.nl” as well as their energy labels. With this, the paper demonstrates how machine-learning algorithms can be used to complete datasets in urban modelling. Using random forest algorithms, Xg boost, support vector machine, multinomial and ordinal logistic regression and an artificial neural network, the thesis finds that only two models - Random Forest and SVMC - performed with reasonable accuracy when predicting the building type, whereas the energy label prediction was unsatisfactory in all models and strategies. The best performing model was achieved with a model strategy where the Random Forest algorithm is trained separately for each town.

Keywords: machine learning, classification, energy label, building type, real estate object, urban planning, topography

Acknowledgements

I would like to thank my thesis supervisor Derek Karssenberg as well as Sjors Hoek and Stijn van der Hoek from GeoGap for providing valuable feedback and help.

Acronyms

NN	Neural Network
ML	Machine Learning
SVMC	Support Vector Machine Classifier
RF	Random Forest
GIS	Geographic Information Systems
BAG	Basisregistratie Adressen en Gebouwen/ Key Register Addresses and Buildings

Chapter 1

Introduction

Machine learning techniques enable companies to approach recommendations and missing data in a new way and make more educated guesses. In the real estate industry, predicting data such as the energy label of a building or its building type can be of value for multiple reasons. First, given the ongoing challenges of climate change, quickly identifying energy-inefficient buildings can play a role in making homes more sustainable considering that housing is a major contributor to greenhouse gasses [2]. Moreover, if you know the building type, you can provide this data to companies that are looking for specific types of buildings to use them, for instance as office space. Knowing the building type will also provide hints about the energy label of a real estate object. The US Department of Energy 2011 report, for example, found that apartments use about half as much energy as freestanding houses [1]. With the energy label itself, companies can better estimate their buildings' energy consumption without requiring a physical evaluation of the buildings to determine their energy classes. Companies that aim to renovate certain buildings can also target their marketing to entities that have buildings in less energy-efficient classes. In addition, private individuals who are looking for energy-efficient residences would benefit from knowing the most likely energy label of their housing object of interest. Therefore, this master thesis aims to find a machine-learning algorithm that can successfully predict the building type and the energy label of a real estate-object. The predicted building type and energy class are intended to be imple-

mented in the databases and the product ”kadastralekaart.com” of the company ”GeoGap”. With this data, the company plans to provide customers of the data platform with an estimated energy label and building type for all residences (mostly referred to in this thesis as real estate objects) in the Netherlands.

1.1 Research question

This thesis aims to answer the research question, “which ML algorithm performs the best at predicting the energy label and building type of a large sample of Dutch real estate objects (BAG)?. As a sub-question, this thesis investigates the features necessary to be included in the algorithms to create a sufficiently accurate model using different model building strategies. Moreover, this dissertation investigates whether one can find differences in classification accuracy between energy label and building type in each of the variable’s categories.

1.2 Major contribution of the Dissertation

Alongside the aforementioned ecological and commercial motivations of this research, this dissertation also contributes to the existing academic literature by seeing if the previous findings on the performance of certain algorithms to predict the building type and energy label can be converted to the dataset of Dutch real estate objects. In addition, this paper is able to make conclusions on the difference between the performance of algorithms on predicting either the building type or the energy label. This comparison has not been made before. Lastly, validating a method to accurately predict missing data such as the building type can be applied to help other scholars in city planning or topography

to describe the ”small-scale description of settlement structures” [3].

1.3 Organization of the Dissertation

The first chapter of this paper begins by introducing the reader to the importance and relevance of my research, both from a commercial and, in a separate part, an academic perspective. After outlining the organization of this paper, the definition and purpose of ML, building types and energy labels are presented. In chapter 2, a literature review of existing publications about the predictions of energy labels and building types is provided. Chapter three presents the methodology of this dissertation, which consists of a presentation of the exploratory data analysis, data wrangling, feature selection, model building strategies and model evaluation part. It also includes a description of the classifying machine learning algorithms, such as random forest or support-vector machines. In Chapter 4, the results of each algorithm and the model building strategy are presented. Chapter 5 discusses the results and its implications. Chapter 6, then, summarizes the dissertation, it’s limitations and provides ideas on future research possibilities that can build upon the findings of this thesis.

1.4 Definition and purpose of machine learning

A computer algorithm that learns from existing data is referred to as machine learning. For the learning process, these algorithms typically use a large amount of data and a small number of input features. There are two forms of learning: unsupervised and supervised machine learning. Unsupervised learning can detect patterns and structure in data without having to learn these structures beforehand. In supervised learning, ML models are trained with an adequate

amount of labelled data in the outcome variable, and can be used to forecast the outcome variable for unlabeled samples without knowing the relationship between the variables. This relationship will automatically be detected by the algorithm. [2] This dissertation use supervised ML techniques since the information the ML algorithms aims to predict from the data is already defined in around 50 percent of the outcome variables energy label and building type. The key advantage of using machine-learning techniques to interpolate missing data in housing is that one can potentially assess the energy label of a real estate object, its building type and its relationship to the input variables without having to physically observe the building and its energy consumption or acquire a large amount of domain knowledge beforehand.

1.5 Definition and purpose of building type

Although there is not a single definition of the different building types – it varies from industry to industry – this thesis will refer to the subsequent classification since these are the building types available in our database of energy labels and building types presented in section 3.1. A visual depiction of each of the building types can be found in the appendix.

- Apartment (other)
- Porch House
- Terraced house between/Row house
- Detached house
- Maisonette
- Gallery-access apartment
- Semi-detached and terraced house corner
- Residential building with non-independent accommodation

This classification provides information about the building itself that potential buyers and tenants seek. Moreover, in the methodology section, the paper will

present that it is also a significant predictor of the energy label itself.

1.6 Definition and purpose of energy classes

Energy labels are a form of energy certification. ”Energy certification is mainly a market mechanism whose main objective is to promote higher energy performance standards than those already regulated.” [4] In January 2003, the European Union introduced the Energy Performance of Buildings Directive (EPBD), which had the stated purpose of encouraging building energy efficiency improvements. As the EU stated: ”Member states shall ensure that, when buildings are constructed, sold or rented out, an energy performance certificate is made available to the owner or by the owner to the prospective buyer or tenant.” [4]. This has led to most of the European Union member-states, including the Netherlands, to adopt the EU energy certificate for buildings. [5].The official energy label ranking by the EU and the ranking used in our dataset has the following order from the lowest to the highest-ranking:

- G
- E
- C
- A
- A++
- A++++
- F
- D
- B
- A+
- A+++

The energy certification evaluation, which should be performed by a licensed technician, involves examining thermal features, building shell, airtightness, thermal equipment, lighting, orientation, renewable energy, ventilation, and internal temperature conditions. [4] This physical evaluation is the most accurate to obtain the energy label. However, being able to interpolate the energy label using

data such as among others the construction year of the residence or the building type can be cost-saving and much faster.

Chapter 2

Literature Review

2.1 Literature on predictions of energy classes

This paper is not the first with the aim of predicting the energy consumption or the energy label of a real estate object using a single or multiple machine -learning algorithms. In one of the more recent publications, Seyedzadeh et al. [2] reviewed four primary machine learning (ML) approaches performed by prior scholars: artificial neural networks, support vector machines, Gaussian-based regressions, and clustering, all of which had already been used in forecasting and optimizing building performance. Among these was for instance a study by Buratti et al [6] which exclusively used neural networks on 6500 energy certificates (2700 are self-declaration) received by the Umbria Region (central Italy), to assess the global energy consumption of buildings from multiple and specific features denoted in the certificates. Buratti et al managed to predict the energy labels with an error rate of only 3.6 percent from all predictions. Seyedzadeh et al. [2] also present a study by Edwards et al [7] which makes a comparison of support vector machine algorithms and artificial neural networks to predict the hourly energy consumption of small residential buildings, resulting in an NN being the least accurate model. After reviewing all the literature on using ML to predict energy consumption, Seyedzadeh et al. [2] conclude that "as from literature, it can be induced that all models provide reasonable accu-

racy by supplying large samples and optimizing the hyper-parameters. Thereby, it is imperative to thoroughly analyze the nature of available or collectable data and the application, to choose the most suitable model." This shows again that our work is valuable, since our data varies from the priorly used databases, and hence a new comparison of ML techniques is worthwhile. Moreover, a comparison of such a large number of ML algorithms to predict energy labels on the same database has not been made before.

2.2 Literature on predictions of building types

Building footprint classification is a relatively young field that is currently in the early stages of development. It has been more important in recent years as digital high-resolution imagery and vector data has become more widely available. Existing methodologies differ greatly in their stated building typology depending on the objective and topic of investigation. [3]. Many of the first studies on the classification of urban structures have for instance been applied in the field of urban modelling. In contrast to this paper and Hecht et al.'s study, these publications chose the technique of remote sensing, which uses a recording device such as an airplane with electromagnetic sensors to obtain information of properties or buildings [8]. In the more recent research on building classification, Hecht et al identified two main different strategies. The first is the top-down or 'knowledge based approach' in which explicit knowledge is taken from expert manuals, rule sets or descriptions with the advantage of high transparency and replicability, albeit with little ability to adapt to new input data, is used for classification. The second bottom-up or 'data-driven approach' is the strategy Hecht et al. and our study uses. In this approach, building type classification is performed by a computer using either an unsupervised (e.g. clustering) or a supervised machine-learning algorithm (e.g. random forest) based on a given training data-set like in our and Hecht et al.'s study.

Hecht et al. also present the already applied supervised building classification approaches from the years 2000 until 2008 [9–13], among which Steiniger et al [11] executed the first comparison of different classification algorithms to classify urban structures. The study that this dissertation focuses the most on was carried out by Hecht et al. in 2018 [3] whose objective was to produce a machine learning approach for automatic classification of building footprints for the small-scale description of settlement structures. Next to the data input comparison, he used 16 different machine learning classifiers (e.g. linear discriminant analysis, artificial neural networks, SVMs, decision trees and ensemble methods). This showed "that building footprints obtained from topographic databases such as digital landscape models, cadastral databases or 3D city models can be classified with an accuracy of 90–95 percent" using a random forest algorithm.

Chapter 3

Methodology

3.1 The available data

There are two datasets that will be used for this analysis. The first is one is the BAG dataset, which compromises 9,236,823 residences in the Netherlands at the time of writing. It can be retrieved from ”kadaster.nl” and contains the address, GIS data (mid-point and layout of the building) of each residence in the Netherlands, as well as features such as surface area, the purpose of the building and its construction year. From the house number I created a categorical valuable that returns a ”1” if the house number is a ”1” and if the house number is larger than ”1”, it returns a ”0”. Since I expect many corner houses to be the first in a street, with this labelling I aim to predict a large part of the buildings with type ”Semi-detached and terraced house corner”. The second data frame contained the target variables: The energy labels and building type of 4,687,182 real estate objects in the Netherlands. It is provided by ”rvo.nl”, the Dutch government’s agency for enterprise. By merging these two data frames, I receive a dataset with a large number of features with one part where the target variables are not missing, which I can use for training and testing, and the rest of the data where the target variables are missing which can be used for predictions. The rest of the data should account for roughly 50 percent of the missingness in the target variables as 4,687,182 rows of the

second data frame divided by 9,236,823 rows of the first data frame equate roughly 0.51.

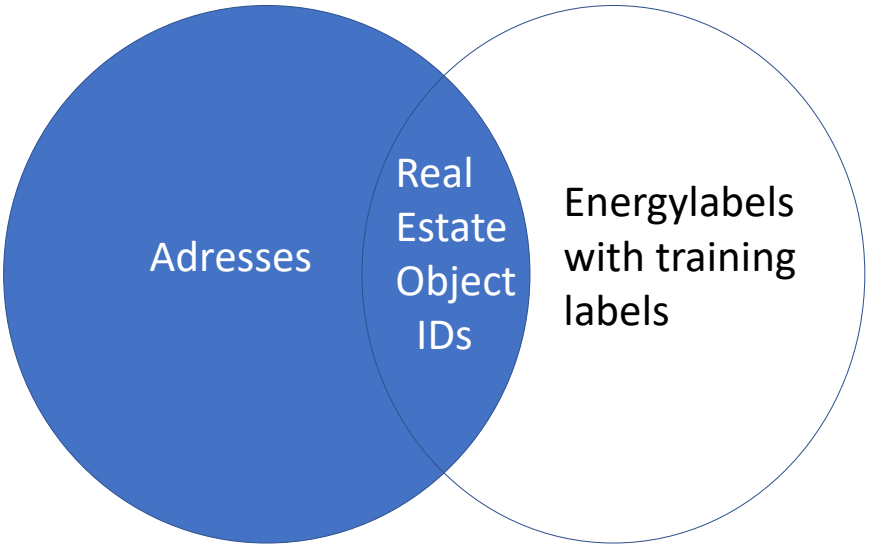

Figure 1: Visualization of the left join

3.2 Exploratory Data Analysis and Merging

The analysis begins by exploring the two datasets that were described in the section 3.1. I used Pandas Profiling to obtain a thorough first understanding of the data, such as how many missing values each data frame has and in which columns. Pandas Profiling is a python package that ”automatically creates data visualizations and summary statistics of data frames” [15]. Knowing the missing values and columns that contain the same value for all rows or IDs that have no predictive value for the analysis, I could already exclude certain features for the later prediction and reduce computation effort. I could then again read in the two complete datasets into my python environment, using only the columns where I did not find too many missing values or that were not relevant for the analysis. Both of the two data frames contained a real estate ID, on which

I could then perform a left join to merge them. In this way, I obtained a data frame with the aim to have enough predictive information to interpolate the target variables energy label and building type.

3.3 Feature Selection

As previously mentioned, certain features could already be excluded since they had too many missing values or, had no relevance or variance. For the rest of the variables, I use the regression feature selection techniques chi-square test and, in the subsequent model fitting part, recursive feature elimination to select variables that are significant in predicting our target variables building type and energy labels. Chi-square is used due to its ease of application and since the majority of my independent variables are categorical. For the continuous variables, I will rely on recursive feature selection, which refers to the act of starting the model with all available features and reducing the number of variables to see if the fit of the model improves [16]. Using these techniques, I end up with the following features, which I will select based on the model building strategy presented in section 3.5.

- city name (categorical)
- province name (categorical)
- building purpose (categorical)
- building type (categorical)
- street name (categorical)
- municipality name (categorical)
- construction year (continuous)
- surface area (continuous)
- house number (categorical)
- postal code (categorical)

3.4 Data Wrangling

To prepare the data as an input for the ML model, it needs to be prepared, so that it can be used as input for the classification models. This is: the data wrangling step. Firstly, I added strings to every cell in the columns that indicated location, which are city, municipality and province. By doing so, I avoided the confusion of cities, provinces and municipalities with the same name. Certain ML algorithms also do not accept duplicate feature names. Secondly, I merged building types that had duplicate categories, such as " Apartment" and " Apartment (other)". Thirdly, I specified and reordered the levels of the target variable " energy label" since they were ordinal and, as previously mentioned, excluded columns of variables with too much missing data or non-predictive potential and split the data into a training dataset, test dataset and a missing data set. Lastly, the data wrangling involved turning the categorical variables into factors using label encoding for the y variable and one-hot encoding for the x variables. One-hot encoding and label encoding are two distinct ways of transforming the categorical data to vectors of real numbers [14]. I use label encoding for the y variable since most of the ML model require one-dimensional inputs for the outcome variable , which label encoding provides as an output.

3.5 Model building strategies

Before applying the algorithms to assess their performance, I outlined four distinct model building strategies to obtain a satisfactory and ideally replicable model. An overview of the strategies is shown below in Figure 2. The first strategy is to use the whole complete-case dataset (no missingness) as training data at once to have a model that benefits from having been trained on a large

amount of data (roughly 4,687,182 rows) and to be able to build a model that can be used for predictions on any real estate object in the Netherlands. In this first strategy, I used all features that indicated location, excluding street name and postal code. If this strategy failed, possibly due to too high a computational burden, a sample of 25,650 objects was used instead. In this second strategy, I made sure that the building type and energy classes that were underrepresented in the dataset were overrepresented in our sample, so that the models could exploit the maximum of number of objects from the minority classes. For the first two strategies, especially the first, the sizes of the files, which in total amount to 14 gigabytes or nine million rows, required me to work in Microsoft Azure, a cloud service where one can store large amounts of data as well as perform python commands on this data [17]. In the third strategy that I adopted, if the previous strategies failed due to an insufficient prediction accuracy or computational burden, I trained the ML algorithms separately on two towns in the Netherlands: Utrecht, to have a larger city as an example, and Cothen, a randomly selected village in the province of Utrecht. As a consequence, I am able to compare the test accuracies between an example village and city. However, I will apply only the algorithm that performed best on the city data on the village dataset, since performance difference will likely be similar for the other algorithms. If this model succeeds for one city or village, this indicates that it will most likely perform similarly well for other cities or villages. Nevertheless, it has to be retrained anew for each place of residence as this model will also consist of unique streets and postcodes as categorical features to obtain more predictive power. The fourth strategy is to include the variables that were included in the previous town-only strategy III on a larger sample scale, using the same sample used in strategy II.

ML building strategy	I Entire dataset	II Sample	III Town only	IV Sample with Street& Postal code
Feature				
province name	x	x		
municipality name	x	x		
town name	x	x		
street name			x	x
house number	x	x	x	x
postal code			x	x
constuction year	x	x	x	x
purpose of use	x	x	x	x
surface area	x	x	x	x
postal code			x	x

Figure 2: Model Building Strategies and their used features

3.6 Predicting building type, algorithms used

The analysis begins by predicting the building type. As a consequence, I can use the successfully building type to predict the energy class in the dataset where the outcome variables are not yet defined once the best model has been found. Predicting the building type requires computer algorithms that can predict category classifications of more than two categories, since the outcome variables building type and energy label are categorical and have more than two categories. The algorithms that I present in the subsequent parts are commonly used in supervised ML for multiple category classification problems [19] and were partly used in previous energy label or building classifications as outlined in section 2.1 and 2.2.

3.6.1 Random Forest

This algorithm was used first in the analysis of this thesis, as Hecht et al. [3] stated that ”The Random Forest (RF) Algorithm was identified as the best learn-

ing algorithm for building classification". The RF Algorithm is an ensemble classifier that uses a large number of trained decision trees to classify data. A random bootstrap selection of the training data is used to create the individual trees.

3.6.2 XGBoost

The XGBoost algorithm is an enhancement of the gradient boosting algorithm, which was developed for very high predictive capability. In XGBoost, individual trees are produced using multiple cores and data is arranged to minimize the search times. As a result, model training time is reduced, and performance is improved. [20]

3.6.3 Multinominal Logistic Regression

Based on numerous independent factors, multinomial logistic regression is used to predict categorical placement or the probability of category membership on a dependent variable. Both dichotomous and continuous independent variables can be used. Multinomial logistic regression is a straightforward expansion of binary logistic regression that allows for more than two dependent or outcome variables. Like binary logistic regression, multinomial logistic regression utilizes maximum likelihood estimation to estimate the probability of a data point belonging to a certain category. [21]

3.6.4 Support Vector Machine

Just like the other algorithms this paper presents, a support vector machine (SVM) is an algorithm that uses supervised learning to assign labels to objects. [22] SVM maps training examples to points in space in order to widen the distance between the two categories as much as possible. New examples are then mapped into the same space and classified according to which side of the gap they fall on. [23]. SVM does not allow multiclass classification natively in its most basic form. Thus, I will use SVM Multiclass Classification, which breaks down a multi-class problem into a binary problem.

3.6.5 Neural Networks

Classification is one of the most active research and application areas of neural networks. Neural networks are a right-brained approach to artificial intelligence that recognize patterns based on previous training and are motivated by the brain. The fully connected three-layer model, which entails an input layer, hidden layer, and output layer, is one of the simplest and most frequent neural network models [24] and is also the model, this paper will make use of for the classification problem at hand.

3.7 Predicting energy classes, algorithm used

3.7.1 Ordinal Logistic Regression

Next to the algorithms already used for predicting the building type, the paper makes use of ordinal logistic regression instead of multinomial logistic regres-

sion since unlike building type, the outcome variable energy class has an order from lowest to highest as seen in section 1.5. Like multinomial logistic regression, ordinal logistic regression is a derivate of binary logistic regression, with the difference that multinomial logistic regression does not consider the ordering of the outcome variable. Therefore, this paper applies ordinal logistic regression using cumulative probabilities, cumulative odds and cumulative logits which takes the ranking of the outcome variable into account. [25].

3.8 Evaluation of accuracy and model selection

To evaluate the performance of each algorithm and each of the four model building strategies, accuracy scores (correct classification divided by the total number of classifications) are used to assess the performance of each algorithm. For the best performing algorithm in terms of computational time and accuracy, confusion matrices are employed. These allow for the identification of outcome variable categories that are harder to predict than others. The evaluation is done by splitting the final datasets of each model's building strategy into a training dataset (70 percent) and a test data set (30 percent). The training set is used to train the model of each algorithm and the test set will be utilized to assess the accuracy of the predictions, as well as to see if there is an indication of overfitting (e.g. if training accuracy is significantly higher than test accuracy). Instead of the simple train/test split, for the neural networks, I use repeated k-fold cross-validation and took the average accuracy of all accuracy scores. In addition, I evaluated the mean test accuracy for each algorithm throughout all model building strategy, and vice versa. Overall, the research aimed to create an ML model capable of forecasting energy class and building type with an accuracy of at least 80 percent, ideally above 90 percent. I, then, utilized the ML algorithms that performed with the highest accuracy above the threshold of 80 percent test accuracy and a reasonable computational time to predict the

unlabeled real estate objects in the entire dataset.

Chapter 4

Results

4.1 Results-Model Building Strategy

4.1.1 Building Type

ML building strategy	I Entire dataset	II Sample	III Utrecht /Cothen only	IV Sample with Street&Postcode	MEAN TEST ACCURACY
Algorithm					
Random Forest	Nan	0.54	0.84 / 0.72	Nan	0.69
Support Vector	Nan	0.26	0.85	Nan	0.555
XG boost	Nan	0.43	0.77	Nan	0.6
Multinominal	Nan	0.26	0.57	Nan	0.415
Neural Network	Nan	0.35	0.3	Nan	0.325
MEAN TEST ACCURACY	Nan	0.368	0.666	Nan	0.517

Figure 3: Accuracy Score Matrix Building Type

When comparing the four different model building strategies I outlined beforehand, it was shown that the best performing model building strategy was to compute a model for each city separately, including the street name and the postal code as categorical dependent variables. I demonstrated this using the example of Utrecht as a larger city, and Cothen (province of Utrecht) as a smaller village. On average, strategy III had a test accuracy of 67 percent throughout

all models, and thus a much higher test accuracy than all other models' average accuracy. Most importantly, the best performing model of 85 percent was the Support Vector Machine Classifier (SVMC) using strategy (III) on the city of Utrecht, with a test accuracy of 85 percent. Strategy I, wherein I tried to work with as much training data as was available, was not successful due to memory usage problems. The same problem was faced when trying to train the ML model using the same sample as in strategy II and including the postal code and street name as features, which was strategy IV. These findings show similarities with Hecht et al.'s [**?**] work, as their study also worked exclusively with models trained on individual cities rather than an entire country. However, this also has to do with the fact that data types differed from city to city in their dataset.

4.1.2 Energy label

The model strategies applied on the energy label prediction showed a similar outcome as for the building type. Again, memory usage problems were inevitable when trying to work with the entire dataset, as in strategies I and IV. Again, the local model (III) performed the best, albeit with much less accuracy than for the building type, with a mean of 56 percent test accuracy throughout all the ML algorithms and a maximum accuracy of 0.68 in the random forest classifier trained on the city of Utrecht.

Figure 4: Accuracy Score Matrix Energy Label

ML building strategy	I Entire dataset	II Sample	III Utrecht / Cothen only	IV Sample with Street &Postcode	MEAN TEST ACCURACY
Algorithm					
Random Forest	Nan	0.56	0.68 / 0.58	Nan	0.62
Support Vector	Nan	0.39	0.46	Nan	0.425
XG boost	Nan	0.48	0.66	Nan	0.57
Ordinal Regression	Nan	0.2	0.54	Nan	0.37
Neural Network	Nan	0.3	0.29	Nan	0.295
MEAN TEST ACCURACY	Nan	0.386	0.526	Nan	0.456

4.2 Results-Comparing the ML algorithms

4.2.1 Building Type

Of all the ML algorithms, the best performing model was the SVMC trained and predicted with the city of Utrecht. With the Support Vector Machine Classifier, a test accuracy of 85 percent was achieved. The next best performer was the Random Forest algorithm, with 84 percent test accuracy. The Random Forest had the advantage of having a much shorter training time than the SVMC model. On average, throughout all strategies, the Random Forest Classifier performed the best, with a mean test accuracy of 0.69 percent. On average, the neural network was the least accurate, with an average test accuracy of 0.295 in all model building strategies.

4.2.2 Energy label

None of our algorithms used reached the target test accuracy of 80 percent when trying to interpolate the energy label. In fact, the best scoring ML algorithm, the Random Forest algorithm, had an accuracy of 68 percent in the local

model and 62 percent overall. This was closely followed by the XG Boost algorithm, with 66 percent locally and 57 percent mean test accuracy. The worst performing ML algorithm was again the Neural Network, with a mean test accuracy of 0.29 percent of strategy II and IV.

4.3 Results-Category Analysis

4.3.1 Building Type

1 Apartment (Other)
2 Gallery-access house
3 Maisonette
4 Porch house
5 Terraced house between
6 Semi-detached / terraced house corner
7 Detached House
8 Residential building with non-independent accommodation

Predicted	1	2	3	4	5	6	7	8
Actual								
1	**10844**	95	324	151	256	18	2	0
2	78	**1593**	4	10	5	0	0	0
3	511	7	**1465**	16	198	16	2	0
4	176	11	18	**1013**	10	4	1	0
5	197	0	52	7	**5050**	466	6	1
6	78	1	17	3	1062	**546**	19	0
7	6	1	0	0	31	40	**91**	0
8	0	0	0	0	0	0	0	**22**

Figure 5: Confusion Matrix Building type Prediction using Random-Forest

With our best performing algorithm, the Random-Forest Classifier, the model resulted in an 84 percent test accuracy of all building types. Figure 3 shows

a confusion matrix where the diagonal line represents the correct classifications for each of the building type categories, from which I can calculate the categories with the highest amount of misclassification relative to the total classifications. This showed that category 5, ”terraced house between”, had the highest number of misclassifications, with 6 percent of all predictions being false. Most of these were due to a confusion with category 6, Semi-detached houses/terraced house corner. The second-highest number of misclassifications come from category 1, Apartment (Other), with 4 percent of false predictions relative to the total number of predictions with misclassifications throughout all categories. Alongside the misclassifications relative to the total number of classifications, I also calculated the misclassifications relative to the number of predictions within one category (also called precision of a model). I found that 50 percent of all predictions within category 6, Semi-detached / terraced house corner, were false, which is the highest number of within-category misclassification among all categories.

4.3.2 Energy label

Predicted	A+	A++	B	C	D	E	F	G
Actual								
A+	**8426**	4	333	137	66	30	16	0
A++	0	**2**	0	0	0	0	0	0
B	337	0	**2072**	437	106	47	17	0
C	128	0	294	**2877**	577	273	97	0
D	35	1	82	574	**1653**	464	144	1
E	11	0	26	244	558	**879**	167	0
E	17	0	29	150	293	343	**169**	0
G	11	0	11	95	209	192	355	**22**

Figure 6: Confusion Matrix Energy Label Prediction using Random-Forest

The random-forest algorithm that was applied to forecast the energy label. It was the best algorithm this thesis found for this task, with a 0.68 percent test accuracy, using strategy III on Utrecht. The largest misclassifications came from

energy classes D and E, with 8 and 7 percent misclassifications from all predictions, respectively. As for the within-category misclassification, the most false classifications appeared in the A++ and F energy class with 71 percent and 82 percent false classifications respectively.

Chapter 5

Discussion

The paper found the random forest model and support vector machine to be the best performing algorithms in terms of test accuracy. This is very much in line with the findings of Hecht et al. [3] who also found the Random-Forest to be the best algorithm at predicting building types and to have the shortest training time. It is also this paper's algorithm of choice due to its balance between accuracy and computational time. Although the support vector machine classified buildings with 1 percent more test accuracy, 9 hours of fitting time, versus 17 minutes for the Random-Forest, is a clear advantage of the RF algorithm, especially considering that models might have to be fitted for each town individually. As for the energy labels, none of our models achieved satisfactory test accuracy. The closest was the Random-Forest algorithm, with 68 percent test accuracy trained on the city of Utrecht. This finding is contrary to Seyedzadeh et al. [2], who showed that "all models provide reasonable accuracy by supplying large samples and optimizing the hyper-parameters". For my model to improve by that much for it to be satisfactory, parameter tuning was, however, unlikely to achieve this, since parameter changes have in my model never led to a test accuracy increase of more than 4 percent. Nor has an increase in the sample size. This indicates that the most feasible way to improve the energy label, and building type classification is to source more features as input variables. Possible features could be based upon the GIS data, which I excluded from my ML models. This opportunity for model improve-

ment will be looked at more closely in the section on future research (6.3). Moreover, this paper revealed that it is much harder to predict the energy label than the building class, given our input data. I suspect again that this could be overcome by adding statistically derived spatial features to the model. This thesis also showed a difference in model accuracy between a small town and a larger city. One major reason could be the settlement structure differences between villages and cities. Cities like Utrecht are characterized by large complexes of social housing of a very similar architectural style, such as in the area of Overvecht [26],], which leads to numerous real estate objects of the type ”Apartment (other)” that can easily be distinguished from other building types. Lastly, this thesis demonstrated a large number of misclassifications between terraced houses and semi-detached/terraced corner houses, which once again highlights the importance of spatial data among the features. Its advantages will be further explained in section 6.3 since the difference between these two categories is nothing but the distance to the closest neighboring house on one side. The misclassifications in our best energy label prediction model did not follow any visible trend and since with increasing or decreasing energy labels the model became more or less accurate, conclusions cannot be made from the results .

Chapter 6

Conclusion

6.1 Limitations

Firstly, a limitation of this study was the lack of features in comparison to previous studies. Whereas this thesis employed only a maximum of seven features, Hecht et al. [3] used between 72 features (data type I) to 87 features (data type V) which could allow for a much more accurate model, albeit a less replicable one if the same features can not be gathered from the database at hand. Secondly, the underrepresentation of certain categories in the dependent variable ”building type”, for instance ” houseboat” or ”log house”, was only represented less than 30 times in our dataset and not present at all in our local strategy III models. Thus, a new study that develops or makes use of a feasible and effective strategy to train the model for such underrepresented categories should be performed to be able to predict this minority group in the building type more accurately. One strategy could be to first sample the minority strategies, as was done in model strategy II, and then build a separate machine learning model for each class in the dependent variable. Lastly, two of our model building strategies failed due to computational limits. This might be overcome with a computer or distributed computer system with more RAM, or a sophisticated strategy that separates the data into chunks and processes it this way.

6.2 Future Research

As discussed, the models built could have been improved in their accuracy by including more features. One of the options available directly from one of the original addresses (BAG) available freely online is spatial statistical GIS data, which was not included due to time constraints. This GIS data includes, for example, the middle point of each real estate object. From this middle point (centroid), one could calculate the distance to the closest neighboring real estate object. This distance can be useful in the prediction of the building type as well as the energy label. The minimum distance from the centroid is, also used by Seyedzadeh et al [2] in one of their unsupervised learning algorithms. It can, for instance, indicate whether the residence is a row house or a free-standing house, since free-standing/detached houses have a larger distance to their neighbors than a row house. Free-standing houses are also less energy efficient than row houses [1]. Hence, I expect the minimum distance to also be a significant predictor of the energy label. From the minimum distance to the closest neighboring object, one can also derive an additional feature that returns the energy label or building type of the closest neighbor. For example, I believe that knowing that the closest real-estate object is a detached-house, for example, drastically increases the likelihood of the actual object also being a detached house, as thise thesis showed that knowing the street a residence is already a significant predictor of the housing object type. Future publications could also further investigate the possibility of a model directly applicable to the entire country, which failed in our country due to memory problems when trying to include all streets as a categorical variable (IV) and accuracy problems when working only with a sample (II). Scholars with more memory available or another strategy to process this large amount of data could hence build upon strategy V. Overall, future, academic research should further broaden the relatively scarce research [3] of automatic classification of building types.

6.3 Conclusion

The objective of this thesis was to compare different ML algorithms to predict a real estate object's energy label and building type. After the data preparation, I applied five different algorithms on each variable using four different model building strategies that essentially differentiated between trying to fit a model that works for the predictions on the entire dataset or only on one town. The model that achieved the highest accuracy (while maintaining a short enough run time) in predicting our target variables was the Random-Forest algorithm, with 84 percent for the building type and 68 percent for the energy label. Thus, our aim of at least 80 percent test accuracy was partially accomplished. The results section revealed, firstly, that it was harder for the algorithm to interpolate energy labels than building types; secondly, that it was harder for the algorithm to interpolate, residences' building types and energy labels in more rural areas than in cities; and, finally, that certain housing types and energy label classes were harder to predict than others. Due to these findings, extending the model with additional features that rely on the GIS data columns provided in the already used data is desirable.

Bibliography

[1] H. Philipsen N. The Energy Footprint of Apartments, Rowhouses and Houses [Internet]. Rge energy footprint of different dwelling types and locations. 2014 [cited 2021Jun10]. Available from: https://network.aia.org/communities/communityhome/digestviewer/viewthread?GroupId=1993

[2] Seyedzadeh S, Rahimian FP, Glesk I, Roper M. Machine learning for estimation of building energy consumption and performance: a review. Visualization in Engineering. 2018 Dec;6(1):1-20.

[3] Hecht R, Meinel G, Buchroithner M. Automatic identification of building types based on topographic databases–a comparison of different data sources. International Journal of Cartography. 2015 Jan 2;1(1):18-31.

[4] Rey FJ, Velasco E, Varela F. Building Energy Analysis (BEA): A methodology to assess building energy labelling. Energy and Buildings. 2007 Jun 1;39(6):709-16.

[5] Kok N, Jennen M. The impact of energy labels and accessibility on office rents. Energy Policy. 2012 Jul 1;46:489-97.

[6] Buratti C, Barbanera M, Palladino D. An original tool for checking energy performance and certification of buildings by means of Artificial Neural Networks. Applied Energy. 2014 May 1;120:125-32.

[7] Edwards RE, New J, Parker LE. Predicting future hourly residential electrical consumption: A machine learning case study. Energy and Buildings. 2012 Jun 1;49:591-603.

[8] Khorram S, Koch FH, van der Wiele CF, Nelson SA. Remote sensing. Springer Science Business Media; 2012 Feb 27.

[9] Sester M. Knowledge acquisition for the automatic interpretation of spatial data. International Journal of Geographical Information Science. 2000 Jan 1;14(1):1-24.

[10] Raheja JL. Recognition of 3D settlement structure for generalization (Doctoral dissertation, Technische Universität München).

[11] Steiniger S, Lange T, Burghardt D, Weibel R. An approach for the classification of urban building structures based on discriminant analysis techniques. Transactions in GIS. 2008 Feb;12(1):31-59.

[12] Römer C, Plümer L. Identifying architectural style in 3d city models with support vector machines. PFG Photogrammetrie, Fernerkundung, Geoinformation. 2010 Nov 1:371-84.

[13] Henn A, Römer C, Gröger G, Plümer L. Automatic classification of building types in 3D city models. GeoInformatica. 2012 Apr;16(2):281-306.

[14] Hancock JT, Khoshgoftaar TM. Survey on categorical data for neural networks. Journal of Big Data. 2020 Dec;7:1-41.

[15] Bantilan N. pandera: Statistical Data Validation of Pandas Dataframes.

[16] Chen XW, Jeong JC. Enhanced recursive feature elimination. InSixth International Conference on Machine Learning and Applications (ICMLA 2007) 2007 Dec 13 (pp. 429-435). IEEE.

[17] Chappell D. Introducing windows azure. Microsoft, Inc, Tech. Rep. 2009 Dec.

[18] Abdul-Rahman A, Pilouk M. Spatial data modelling for 3D GIS. Springer Science Business Media; 2007 Sep 23.

[19] Osisanwo FY, Akinsola JE, Awodele O, Hinmikaiye JO, Olakanmi O, Akinjobi J. Supervised machine learning algorithms: classification and comparison. International Journal of Computer Trends and Technology (IJCTT). 2017 Jun;48(3):128-38.

[20] Ramraj S, Uzir N, Sunil R, Banerjee S. Experimenting XGBoost algorithm for prediction and classification of different datasets. International Journal of Control Theory and Applications. 2016;9:651-62.

[21] Starkweather J, Moske AK. Multinomial logistic regression.

[22] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. InProceedings of the fifth annual workshop on Computational learning theory 1992 Jul 1 (pp. 144-152).

[23] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press; 2000 Mar 23.

[24] Snyder RM. Neural Networks for the Beginner.

[25] Bender R, Grouven U. Ordinal logistic regression in medical research. Journal of the Royal College of physicians of London. 1997 Sep;31(5):546.

[26] Koster M. Citizenship agendas, urban governance and social housing in the Netherlands: An assemblage approach. Citizenship Studies. 2015 Feb 17;19(2):214-28.

Appendix A

Building Type Description

Apartment/Flatwoning

Gallery-access Apartment

Maisonette

Porch House/Portiekwoning

Terraced House/Row

Semi-attached corner house

Detached House

Residential building with non-independent accommodation