Energetic Restoration Pressure
Master Thesis

Lucerne University of Applied Science and Arts
Master of Science in Applied Information and Data Science (MScIDS)
Autumn Semester 2022

handed in by

Sarah Schneeberger
Zähringerstrasse 49
CH-3012 Bern
sarah.schneeberger@stud.hslu.ch

Berne, 22 December 2022

Supervisors

Prof. Dr. Philipp Schütz
Lucerne School of Engineering and Architecture
Technikumstrasse 21
CH-6048 Horw
philipp.schuetz@hslu.ch

Dr. Esther Linder
Lucerne School of Engineering and Architecture
Technikumstrasse 21
CH-6048 Horw
esther.linder@hslu.ch

Co-Lecturer

Thilo Weber
geoimpact AG
Heinrichstrasse 267
CH-8005 Zürich
thilo.weber@geoimpact.ch

Management Summary

Global warming is happening, and its consequences, such as melting Arctic sea ice, rising sea levels, and more severe wildfires, are being felt worldwide. Action to reduce greenhouse gas emissions is urgently needed. In Switzerland, the building sector is one of the main contributors to the country's greenhouse gas emissions and therefore plays a significant role in the country's climate policy. In the long term, this sector must become free from carbon dioxide emissions, which can be achieved through renovations that reduce energy consumption and incorporate renewable energy sources. However, stakeholders such as communities, products and service providers, and homeowners need information on the buildings to prioritize and plan renovations. Geoimpact AG makes an important contribution to this with the platform Swiss Energy Planning (SEP), where general information about the building and the renovation pressure is provided. With the renovation pressure, buildings with a high probability of a near-future investment can be identified.

This thesis contributes to the differentiated consideration of renovation pressure by estimating an energetic restoration pressure. Energetic restorations are renovations of the building envelope that can reduce an existing building's carbon dioxide emissions or energy consumption. In particular, the aim of this thesis is to determine which approach achieves the best performance in predicting the energetic restoration of a building. Detailed information on renovation activities is needed to model an energetic restoration pressure, which is extracted from building applications. Thereby the thesis makes an essential contribution to examining building applications' potential as a source of information for renovations. Until today, the only way to get reliable and detailed data on renovations is to conduct surveys, which is very time-consuming.

The study identified several challenges in classifying building applications, including identifying renovations, misspellings in the description, and the short length of the description. Concerning these challenges, rule-based systems based on preexisting categories and descriptions have been used to classify energetic restorations. The classification was evaluated with a heat energy demand model, which estimates the energy demand, and with a sample of renovation information from an online survey conducted by TEP Energy. According to the heat energy demand model, energetic restorations show a rather low energy-saving potential. However, since reference classes also tend to have low potential, the model tends to underestimate the energy savings potential. This can be caused by the quality of the underlying data, the quality of the classification, or the aggregation of different measures into one class. However, a lack of data quality was also detected in comparing the building application classification with renovation activities identified through the survey of TEP Energy. For more than 60% of the sample, no building application is found. In particular, the classification of energetic restorations only achieves an accuracy of 60%.

To estimate the energetic restoration pressure, logistic regression and the Cox proportional hazard (CPH) model were used. Due to imbalanced data, under-sampling methods were also tried. Different logistic regression models were trained on balanced and imbalanced data. Furthermore, a baseline model using only a few predictors was applied. The logistic regression models were compared with the ROC AUC and confusion matrix. Finally, they were compared to the CPH model by (mean-) ROC AUC. All models achieve very similar results. The CPH model performs slightly worse than the two logistic regressions with under-sampling, which achieve the best ROC AUC of 0.70. However, since the CPH model estimates a time-dependent restoration pressure, and the difference in the evaluation metric is negligible, the CPH is considered the better approach.

Based on these findings, the study concludes that building applications are not a viable source of information for renovations, particularly energetic restorations. Whether they are suitable for certain renovation activities would have to be investigated in more detail by examining which activities require building permits. However, calculating an energetic restoration pressure based on this data is not recommended. It is also advised to investigate the approaches presented in this work on reliable data again. Nevertheless, the CPH model is believed to be a suitable approach to predict the pressure of renovation activities. And it would be interesting to examine and compare more complex models, like Cox's time-varying proportional hazard model and Survival Support Vector Machine, as well as recurrent event methods.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AUC area under the (ROC) curve
BoW Bag of Words
CPH Cox Proportional Hazard
CRISP-DM Cross-Industry Standard Process for Data Mining
EGID Federal Building Identifier
FN false negatives
FNR false negative rate
FP false positives
FPR false positive rate
GBR Gradient Boosting Regression
RBD Federal Register of Buildings and Dwellings (Eidgenössisches Gebäude- und Wohnungsregister, GWR)
HEC heat energy consumption
HEPI Heat Energy Performance Indicator
MAPE mean absolute percentage error
MDAPE median absolute percentage error
ROC receiver operating characteristic
SEP Swiss Energy Planning
SVR Support Vector Regression
TN true negatives
TNR true negative rate
TP true positives
TPR true positive rate

List of Data Sources

Data Category	***Source***	***Extraction Date***
Building Applications	Docu Media Schweiz GmbH	July, 11th 2022
Construction Projects	Federal Register of Buildings and Dwellings (RBD)	Jan., 25th 2022
Building Characteristics	Federal Register of Buildings and Dwellings (RBD)	Sept., 15th 2022
Building Characteristics	Minergie Schweiz	Apr., 15th 2020
Calculations of Geoimpact AG	Geoimpact AG	Sept., 15th 2022
Heat Energy Consumption Biel	Energie Service Biel (ESB)	Sept., 7th 2021
Heat Energy Consumption Geneva	Système d'information du territoire à Genève (SITG)	June, 5th 2021
Heat Energy Consumption St. Gallen	Umwelt und Energie, Stadt St. Gallen	June, 18th 2021

1. Introduction

"Global warming isn't a prediction. It is happening." (Hansen, 2012)

Since the late 19th century, Earth's global average temperature has increased by about 1.1 degrees Celsius. Human activities that have increased carbon dioxide emissions and other greenhouse gases into the atmosphere primarily cause this long-term global warming trend. Consequences of this global warming include: Arctic sea ice is melting, sea levels are rising, and wildfires are becoming more severe ("Earth observatory", 2021). To minimize the extent of these and other consequences, the Paris Agreement requires all United Nations to reduce greenhouse gas emissions (Federal Office for the Environment FOEN, 2018). Under this agreement, Switzerland has also committed to the international climate target: by 2030, it aims to reduce its greenhouse gas emissions by at least 50 percent compared to 1990 levels (Federal Office for the Environment FOEN, 2021).

In Switzerland, the building sector accounts for around a quarter of greenhouse gas emissions and therefore plays an important role in Switzerland's climate policy (Federal Office for the Environment FOEN, 2022). In the long term, this sector should become free from carbon dioxide emissions (Federal Office for the Environment FOEN, 2020). A renovation can massively reduce the energy consumption of a building, and by exchanging a fossil heating system with renewable energies and using an appropriate mix of power sources, emissions of carbon dioxide can be reduced to almost zero (Swiss Federal Office of Energy SFOE, 2022). According to the association of Swiss building envelope companies (2010), 1.5 million buildings should be renovated. For such buildings, the energy savings potential for the building envelope is in the order of 65%. And by replacing windows and renovating roofs, the heat demand can be reduced by 20 - 30% (Federal Laboratory for Materials Testing and Research, 2021). Renovations are therefore necessary for Switzerland to achieve its ambitious energy and climate policy goals. However, only one percent of this building stock is renovated each year, and thus it would take 100 years to renovate all buildings in the country, which would be too slow to achieve these goals (Federal Laboratory for Materials Testing and Research, 2021). Further measures are therefore needed to increase the renovation rate. One approach is to create transparency and identify suitable retrofitting targets. Geoimpact AG makes an important contribution to this with the platform Swiss Energy Planning (SEP). On this platform, daily updated information from building and infrastructure data of different sources is presented (geoimpact AG, n.d.-b). In particular, a renovation pressure is calculated for every building in Switzerland. A high value of renovation pressure implies a high probability of near-future investment in the building (geoimpact AG, n.d.-a). This allows for the identification of possible renovation activities and changes in the Swiss building stock in advance. Weber (2019) states: "Due to the renovation pressure ...

- communities or cantons will quickly find buildings where their initiatives for structural neighborhood planning will fall on particularly fertile ground
- energy suppliers get a good overview of which neighborhoods they should develop first when planning new network infrastructure
- providers of products and services can see directly which buildings have a high chance of success for acquisition
- homeowners receive a well-founded portfolio benchmarking, for example, to create a restoration plan."

This thesis is dedicated to the topic of modeling a specific renovation pressure, more precisely, an energetic restoration[1] pressure. And thereby makes an essential contribution to the differentiated consideration of the renovation pressure.

[1]The term *energetic restoration* will be used solely when referring to structural measures on building envelopes that can reduce emissions of carbon dioxide or energy consumption in the operation of an existing building.

In particular, this thesis aims to examine different approaches to estimate the energetic restoration pressure and to answer the following question:

? *Which approach achieves the best performance in predicting energetic restoration of a building?*

This allows us to identify the buildings in Switzerland with the highest energetic restoration pressure, that is, the probability of a near-future investment in energetic restoration. Thereby, the different customers of geoimpact AG could receive more detailed information in the future and define more targeted measures.

But to be able to answer this question, further questions must first be studied. In particular, information about renovations, particularly energetic restorations, is necessary. Unfortunately, there is no public data source that provides detailed and structured information on renovations carried out for each building. In particular, information on specific structural measures is often missing. Therefore, I try to extract and structure precisely this information from building applications from Docu Media Schweiz GmbH, to finally classify the building applications into energetic and non-energetic restoration. In doing so, the following question must be examined:

(1) *How can building applications optimally be classified into energetic and non-energetic restorations?*

By evaluating these classifications, an essential contribution is made to examining building applications' potential as a source of information for renovations. Until today, the only way to get reliable and detailed information on renovations is to conduct surveys, which is very time-consuming.
Furthermore, this question ensures that labeled data, that is, (non)-energetic restored buildings, are available for modeling the restoration pressure. More precisely, these data form the dependent variable "energetically restored" (yes/no) in the energetic restoration models to be developed. And building characteristics are used as independent variables of the model. When investigating different models for estimating the energetic restoration pressure, the following two sub-questions need to be considered:

(2) *How can energetic restoration be predicted as a time-dependent process?*

The time dependence (in how many years the restoration will be performed) of the dependent variable must be taken into account when creating the energetic restoration pressure models.

(3) *Which metric(s) are most appropriate to measure the performance of energetic restoration models?*

This question is related to validating the energetic restoration pressure models and ensures that the different models can be compared so that the best one can finally be chosen.

But before I examine the research questions in more detail, an analysis of the work of others relevant to the topic is given in Chapter 2. Followed by the terminology used in this thesis. In Chapter 3, I describe the data and the study design. I first look at the classification of the building applications, how the data was prepared, and which models were used. The description of the implementation of the energetic restoration pressure modeling follows this. The results obtained in this study are presented in Chapter 4 and discussed in Chapter 5. In doing so, I also answer the research questions and draw a conclusion.

2. Background

Chapter 2.1 provides a literature review about modeling renovation strategies and predicting renovation probabilities. Chapter 2.2 introduces the terminology used in this thesis.

2.1. Literature Review.

The building sector accounts for a significant share of greenhouse gas emissions and thus is essential for counteracting the climate crisis. Therefore, this topic has been widely analyzed and discussed in research in recent years. Several studies have estimated the potential for reducing greenhouse gas emissions through different energetic renovation strategies in various national building stocks (Mata et al., 2012, 2018; Narula et al., 2018; Polly et al., 2011; Streicher et al., 2019). These studies neglect the time required to implement the strategies. Streicher et al. (2021) estimates that it will take at least several decades to implement such deep energetic renovations of the entire building stock. And during this time, parameters underlying the models change (like the structure of the building stock). To counteract this, he has analyzed 1.1 million retrofit pathways in the Swiss residential building stock considering stock dynamics and climate change impacts (Streicher et al., 2021). They found that the energetic optimal scenario can achieve a reduction of greenhouse gas emission of 90% till 2060, with estimated costs of 9 billion CHF/year. This associated strategy includes early and complete retrofit with the highest energy standard in terms of building technology and building envelope.
If, on the other hand, only a deep building envelope renovation is done, Streicher et al. (2019) has calculated with the Swiss residential building stock model, taking into account respective local conditions and the features of each building, that a reduction of the final energy demand of 57% would be possible in the Swiss building sector.
When considering the energy-saving potential of individual structural measures of building envelope renovation, the insulation of the outer wall has the highest saving potential, followed by the windows replacement (Streicher et al., 2017). About half fewer savings are achieved by insulation of the outer roof and the interior insulation of the floors.
But as Galimshina et al. (2020) shows based on three case studies and by applying the presented framework for robust assessment of renovation strategies in terms of the environmental and economic performance of the building's life cycle, a combination of building technology and building envelope renovation is most efficient. More precisely, the replacement of the heating system should be prioritized, followed by exterior wall insulation and window replacement.

Besides the numerous studies on the potential of different strategies and structural measures to reduce greenhouse gas emissions, the aging process of buildings and building components has also been widely analyzed. The aging behavior is, on the one hand, interesting for modeling future renovation rates; on the other hand, it has a significant influence on renovation strategies. Sandberg et al. (2016) has used a dynamic housing stock model to describe the long-term development of the size and age composition of the housing stock of 11 European countries and to estimate future renovation activities. In all 11 countries, the simulation results in a similar and sobering forecast: the renovation rate increases only between 0.6-1.6%, which is significantly below the 2.5-3.0% renovation rate assumed in many decarbonization scenarios.
In the literature, methods of survival analysis are often used to model the aging process of building components. The name survival analysis originates from clinical research, where they developed methods to estimate the expected lifetime (Pandey, 2020). A key element in survival analysis is the so-called survival function, which gives us the probability that the event has not occurred by the time t. Nowogońska (2019) has estimated this survival function of individual building components with the Rayleigh distribution. From this, they specified the changes in the performance characteristics of a building as the sum of the weighted survival function of individual building components. Also, in the DUREE Project of the Swiss Federal Office for Energy (2019) a method of survival analyses have been used to investigate the lifetimes of four different building elements (Windows, heating systems, facade, and roof). They used survey data on renovations to estimate the survival probability with the Kaplan-Meier estimator. Furthermore, they performed a comprehensive Swiss and international literature analysis to identify

the lifetime of the components based on the state of the art. According to the Kaplan-Meier estimation, the replacement probability of 50% is reached after approximately 30 years for the heating system, 40 years for windows, and between 50 and 60 years for facades and roofs. For the building envelope elements, these numbers are approximately the same as the corresponding predictions based on the literature lifetimes. When moving away from this midpoint, the literature lifetime systematically under-estimate replacement rates among younger building elements and over-estimate replacement rates among older ones.
There are also more complex methods to estimate the survival function, in particular models that include features of an individual/object and their impact on the survival function. Volland et al. (2020) made use of such an approach and applied the Cox Proportional Hazard (CPH) model to relate the time of renovation to a broad set of characteristics identified as important for renovation decisions. In particular, they used data about replacement patterns of building elements from a household survey in Switzerland and estimated the renovation probability of four energy-relevant building elements (heating systems, windows, facades, and roofs). They found no statistically significant relationship between replacement rates and current household income. However, a partially-significant relationship between risk attitude and renovation investment was identified. Other factors that showed a significant effect on the renovation behavior were building type, location, and year of construction classes (before and after 1900). Furthermore, potential renovation delay in specific building groups, such as single-family homes and relatively new buildings, was identified, as well as in the building elements facade and roof.

As we have seen, previous research on modeling renovation probabilities has mainly used information collected from surveys or technically defined lifetime documented in the literature. Conducting surveys is a very time-consuming and costly method. And, technical lifetimes identified by literature reviews are not very accurate in representing the actual renovation behavior (Swiss Federal Office for Energy SFOE, 2019, Volland et al., 2020). In this thesis, I will examine the potential of building applications as a source of information for renovations. The aim of the thesis is finally to use these data to model an energetic restoration pressure. Thereby, I apply and compare logistic regression and the CPH model.

2.2. **Terminology.**

The federal and cantonal building program promotes structural measures to reduce carbon dioxide emissions or energy consumption. The program distinguishes between the following structural measures (Swiss Federal Office of Energy SFOE, n.d.):

- Thermal insulation of the building envelope
- Replacement of fossil or resistance heating systems by heating systems with renewable energies or by connection to a district heating
- Comprehensive energetic renovations or renovations in larger stages as well as new construction according to Minergie-P standard

In the thesis, I will mainly focus on the thermal insulation of the building envelope. In particular, measures to reduce energy consumption for buildings' internal temperature (heating and cooling) are of interest. I explicitly exclude the replacement of heating and hot water preparation systems. Since there is no consistent terminology in the literature for this differentiation, I introduce the following definitions for this thesis:

The term *energetic restoration* will be used solely when referring to structural measures on building envelopes that can reduce emissions of carbon dioxide or energy consumption in the operation of an existing building.

The term *energetic renovation* is used here to refer to all structural measures that can reduce emissions of carbon dioxide or energy consumption in the operation of an existing building. Energetic restoration is thus a part of energetic renovations, as illustrated in Figure 1. For completeness, the term *renovation* is used here to refer to the set of all structural measures related

to buildings.

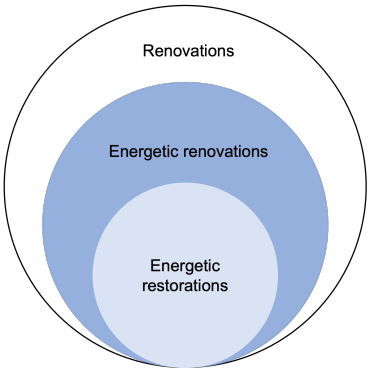


FIGURE 1. Venn Diagram Showing the Relation Between Energetic Restorations, Energetic Renovations, and Renovations

A literature review was conducted and discussed with the business partner to identify concrete structural measures of energetic renovations and especially of energetic restorations. Table 1 contains words related to structural measures, which can reduce emissions of carbon dioxide or energy consumption in the operation of an existing building (Ebert Stoll, n.d.; Gebäudehülle Schweiz, 2010; Streicher et al., 2017; Swiss Federal Office of Energy SFOE, 2022). Thus, with the term energetic restoration, I refer to measures of the group Building Envelope.

TABLE 1. Words Related to Energetic Structural Measures

Structural measures		
Group	***Sub-group***	***Word***
Building envelope	Insulation	Thermal insulation
		Facade insulation
		Vacuum-Insulation-Panel (VIP)
		Compact facade
		Ventilated facade
		Elimination of thermal bridge
		Insulation of basement ceilings
		Screed floor insulation
		Roof insulation
		Removal of the chimney
		Roof greening
	Window	Window replacement
		Triple glazing
		Insulation glass
		Sun protection device (heat insulation in summer)
Building technology	Heat and hot water	Heat pump
		Borehole exchanger
		Geothermics
		District heating
		Wood heating
		Woodchip heating
		Pellet heating
	Solar (heat and power)	Solar heat
		Solar collector
		Photovoltaics
	Ventilation and air conditioning	Comfort ventilation
		Cascade ventilation
		Compound ventilation
		Geothermal cooling
		Free cooling

3. Methodology & Materials

Chapter 3.1 provides an overview of the data used in this study. The study design, in particular, the data preparation steps, the approaches to classify the building applications, and the models to estimate the energetic restoration pressure are described in Chapter 3.2.

3.1. Data Description.

For this study, I used several data sources. Table 2 provides an overview of the available data. The construction projects from Federal Register of Buildings and Dwellings (RBD) were finally not used in this thesis, so I do not discuss this data further in this chapter. Nevertheless, a short overview and analysis of this data can be found in Appendix B.1.

Table 2. Overview of Available Data

Data	*Source*
Building Application	Docu Media Schweiz GmbH
Construction Projects	RBD
Heat Energy Consumption	Energy supplier
Building characteristics	RBD, Minergie Schweiz and geoimpact AG

The heat energy consumption data is based on measurements of heat energy consumption (HEC) of 21'210 buildings in Geneva, St. Gallen, and Biel. The HEC is determined from meter readings from oil, gas, and district heating systems and provided by the energy supplier. Geoimpact AG further processed this data and provided the HEC $[kWh/year]$ of specific years for this work. An important key figure is the Heat Energy Performance Indicator (HEPI) $[kWh/m^2/year]$, which can be derived by dividing the HEC through the energy reference area. In Appendix A (Table A.1), an overview of the provided data is given.

The data from Docu Media Schweiz GmbH has already been preprocessed by geoimpact AG and contains information on 1'519'010 building applications recorded since 2004. Attributes containing free text (written in German, French or Italian), construction costs, existing categorizations, and building information are included. The Federal Building Identifier (EGID) was added via the address information by geoimpact AG and is used to link the data with the other available data. A structured overview is given in Appendix A (Table A.2).

The data on building characteristics provided by geoimpact AG contains information on 2'846'015 buildings. Buildings, which are recorded in the RBD, form the basis of this data set. Geoimpact AG has enriched this data with historized RBD data and self-calculated features. Additionally, the information on whether the building complies with the *Minergie Standard* was added from Minergie Schweiz. Table A.3 in Appendix A lists all available attributes. More information about the calculations from geoimpact AG is described in a post on the SEP platform (Weber, 2022).

3.2. Methodology.

To answer the research questions, a suitable study design is necessary. The Cross-Industry Standard Process for Data Mining (CRISP-DM) is considered the guideline for data science projects and was initially presented by Wirth and Hipp (2000). It consists of 6 steps, which help to plan, organize and implement the project:

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation

– Deployment

The sequence of the steps is not strict, and switching back and forth between different phases is usually required. These interactions are illustrated in Figure 2. In this thesis, I have used this method for the classification of building applications and the modeling of the energetic restoration pressure. The study of the industry partner's problem, the elaboration of the research questions, and the gathering of domain-specific knowledge, which belong to the step "Business Understanding," are described in Chapters 1 and 2. The details of the other individual steps are described in Chapters 3.2.1 and 3.2.2.

FIGURE 2. Phases of the CRISP-DM ("Cross-industry standard process for data mining", 2022)

3.2.1. *Classification of Building Applications.*

Data understanding

Two data sources were available to identify energetic restorations: building applications from Docu Media Schweiz GmbH and data on construction projects from RBD. There are much more building applications in the data of Docu Media Schweiz GmbH than in the RBD. Moreover, the business partner is mainly interested in the potential of these building applications. Therefore, the work is limited to the Docu Media Schweiz GmbH dataset. Nevertheless, the data from the two sources were compared, and the result is described in Appendix B.2. To evaluate the classification, data on measured HEC of individual buildings in the cities of Biel, Geneva, and St. Gallen, and building characteristics from RBD, Minergie Schweiz and geoimpact AG were available. I analyzed these data and checked their quality. The findings are described in Chapter 4.1.

Data preparation

One important goal of the building application preparation was the identification of renovations. Various rule-based logic were tested to assign the building application to at most one of the newly defined application types (new construction, renovation, demolition). Further preexisting categorizations of the building applications were used to identify and delete irrelevant building applications. The finally used rules are described in Appendix C.1 and C.2. And the resulting distribution of identified renovations and new constructions is given in Table 11.

The available free text information of the building applications was merged and tokenized. Lemmatization that is, reducing the words to their dictionary form, was also tested.[2] To apply machine learning algorithms and to perform meaningful analytics on text data, the text content needs to be transformed into numerical feature vectors (Bengfort et al., 2018). Bag of Words (BoW) representation is one of the most intuitive ways to do so, and it represents a text using its word frequency (Pedregosa et al., 2011e). Thereby a word dictionary from the text is created. And the word frequency is stored in a document-term-matrix, where the rows represent the documents, and the columns represent the words of the dictionary. A value (i, j) in the matrix corresponds to the number of occurrences of word j in document i. I have used the BoW representation to find words containing specific word fragments efficiently and analyze the building applications' descriptions. In doing so, the number of occurrences of words related to energetic renovation activities was detected (see Chapter 4.1). Furthermore, the most frequent N-grams (sequence of N words) were identified (see Appendix B.3).

Modeling

I applied rule-based systems to classify the building application by using preexisting categories, providing encoded renovation activities, and by using the preprocessed free text.
Rule-based systems are the simplest form of artificial intelligence. A rule-based system consists of a set of facts and IF-THEN rules that specify how to respond to the set of facts. The knowledge of a human expert can be used to build such rule-based systems (Grosan and Abraham, 2011).

In one rule-based system, preexisting categories were used as facts, and the rules were created with the help of specialist literature and the business partner. More precisely, the categories were assigned to a building element (Group and Sub-group) given in Table 3. Moreover, they were divided into energetic and possible energetic. These assignments were finally used to classify the building applications. In particular, building applications of the application type renovation, containing categories identified as energetic and belonging to the group building envelope, were classified as energetic restoration. The exact rules are described in Appendix D.1.

TABLE 3. Structured Overview of Building Elements

Building elements	
Group	***Sub-group***
Building envelope	Window
	Roof
	Insulation
	Facade
Building technology	Ventilation and air conditioning
	Heating and hot water
	Solar heat
	Solar power

The other rule-based system uses the prepared German text data as facts and the rules were created with the help of professional literature and discussed with the business partner. Words related to energetic structural measures, that is, structural measures that reduce emissions of carbon dioxide or energy consumption in the operation of an existing building, were identified, and a dictionary was compiled (see Table 1 in Chapter 2.2). This dictionary formed the basis for creating a specific but comprehensive dictionary from the text data of the building applications. In particular, word parts of the identified terms given in Table 1 were used to search for corresponding words in the text data of the building application.[3] The resulting dictionary (compare

[2]The analysis of the lemmatized text showed that the lemmatization worked very poorly, and therefore I did not use the lemmatized text further. I suspect this is due to the domain-specific language and the many spelling errors in the building applications. Thus, an algorithm for spelling correction was tested. But it did not work well because it did not recognize the domain-specific words.

[3]This procedure was limited to terms that consist of only one word in German. Therefore, certain construction measures from Table 1, such as removal of chimney and elimination of thermal bridge, are not included in the final dictionary. For simplicity, the insulation of various building elements has been grouped under insulation. However, by using word parts, many spelling errors could be identified and were included in the final dictionary.

Table D.3 in Appendix D.2) was finally used to classify the building applications with the text data. The exact procedure and rules are described in Appendix D.2.

Evaluation

The classification of energetic restoration based on preexisting categories and categories identified as possible energetic were evaluated with the help of the heat energy demand model provided by geoimpact AG. In this model, the HEC is predicted by either using the machine learning technique linear Support Vector Regression (SVR) or Gradient Boosting Regression (GBR). Measurements with a HEPI below 10 $kWh/m^2/year$ and above 300 $kWh/m^2/year$ were identified as outliers and deleted (Gubser, 2021). As in the original model, the predictors were selected and processed (compare Table 4). Finally, the data was split into train and test sets, and a model was trained for each category to be evaluated. Using the test data, the influence of the individual predictors on the HEPI was determined with a partial dependency plot. This means, for each sample and each predictor, the dependent variable is predicted for each possible value of the predictor by leaving the other predictors unchanged. For a boolean predictor, we receive two predictions, which can be used to calculate the HEPI difference. In addition, the mean absolute percentage error (MAPE) and median absolute percentage error (MDAPE) were used to measure the model performance. The results of these evaluations are discussed in Chapter 4.3.1.

TABLE 4. Cleaning and Transformation Steps of the Heat Energy Demand Model

Attribute	***Handling missing values***	***Transformation***
geb_kategorie_gwr	Filling missing values with 0	Encoded as one-hot numeric array
geb_klasse	Filling missing values with 0	Encoded as one-hot numeric array
baujahr_upper	Filling missing values with the median	Standardized and spline transformed
volume_egid	Filling missing values and values < 1 with $3 * energy\_reference\_area\_pred$	Log and spline transformed
energy_reference_area_pred	No missing values	Log and spline transformed
heating_degree_days_pred	No missing values	Log transformed
anz_wohnungen_egid	No missing values	Standardized and spline transformed
heat_meas_incl_hotwater	Filling missing values with 1	Standardized and spline transformed
Dummy variable of the class to be evaluated	Filling missing values with False	Encoded as one-hot numeric array
q_use_true	Deleting building with missing q_use_true	Log transformed

The classification using the text data could not be evaluated with the heat energy demand model provided by geoimpact AG, because the heat energy consumption data were mainly available from the French-speaking area, and the classification using the text data was based on building applications in the German-speaking area. Therefore, the classification based on preexisting categories was used to evaluate the classification with the text data. For this purpose, confusion matrices were created from comparable classes. Confusion matrix is a widespread approach that shows the predicted and actual classification (Düntsch and Gediga, 2019). More precisely, the results of the classification are divided into four groups: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). In our context, the FP group represents all samples that were classified by the text data as members of the class, and by classification based on categories, these samples were classified as not belonging to the class. A few examples are discussed in Chapter 4.3.2. And the complete evaluation is available in the Appendix E.2.

According to the evaluation of the two rule-based methods, the final classification is a combination of the two rule-based systems enriched with categories identified as possible energetic that showed an energy-saving potential in the evaluation with the heat energy demand model. TEP Energy evaluated this final energetic restoration classification by comparing it with 1'000 renovation activities from 2006 to 2019, which were collected in an online survey. To compare

other energetic restoration activities, some new classes (window replacement, energetic facade renovation, and energetic roof renovation) were created. The rule-based logic of these classes is described in Appendix D.3. The results are presented in Chapter 4.3.3. More information about the online survey and the determination of the renovation activities from this survey is described in the project report *Kantonale Energiekennzahlen und CO_2-Emissionen im Gebäudebereich* (Jakob et al., 2021, pp. 33-36).

Deployment
The last step consists of developing and documenting a plan for deploying the model, as well as for monitoring and maintenance. These steps are omitted since deploying the final model is not part of the thesis.

3.2.2. *Modeling of the Restoration Pressure.*

Data understanding
Building characteristics and information if and when an energetic restoration took place are needed to calculate an energetic restoration pressure. Since the same building characteristics were used in the classification step, I already analyzed and checked their quality. The information about energetic restoration results from the classification described in Chapter 3.2.1.

Data preparation
The building characteristics data include all buildings in Switzerland that are recorded in the RBD. Since the restoration pressure is only of interest to existing buildings, the data was reduced to buildings with the corresponding status. To these data, the classified renovation information from building applications was added.
As a further data preparation step, new features were created, like construction cost per building volume and if renovation took place before 2004. How missing values and impossible values were treated is described in Table 5.

TABLE 5. Data Cleaning for Modeling the Restoration Pressure

Attribute	***Handling missing values***	***Handling impossible values***
geb_klasse	Filling missing values with 0	No impossible values
baujahr_upper	Deleting samples with missing values	No impossible values
energy_reference_area_pred	Deleting samples with missing values or values equal to 0	No impossible values
volume_egid	Filling missing values and values < 1 with $3 * energy\_reference\_area\_pred$	No impossible values
anz_geschosse	Filling missing values with $volume\_egid/(3 * geb\_flaeche)$	Deleting samples wiht values > 41
anz_firmen_egid	Filling missing values with 0	No impossible values
projectdate	No missing values	Deleting samples with projectdate before 2004
renovrate_neigh_current	Deleting samples with missing values	No impossible values
renovrate_neigh_event	Deleting restoration with missing values	No impossible values
baukosten/gebäudevolumen	Filling missing values with median	No impossible values
previous_renovation	Filling missing values with False	No impossible values

Depending on the approach, the data were further processed in different ways. One entry was required per building in the first approach, which calculates a time-independent energetic restoration pressure. More precisely, if available, the last identified energetic restoration was selected, and the renovation costs per building volume incurred before the selected energetic restoration were summed. Thereby, the costs of renovations identified as energetic restorations were excluded.
The second approach takes the time dependence of the target variable into account. Therefore, all

energetic restorations were kept in the data set and all buildings without an identified energetic restoration. Per building, energetic restorations within two years were combined and considered one restoration. As in the other approach, the earlier incurred construction cost per building volume was calculated. Furthermore, the lifetime of a building, in particular of an energetic restoration, was calculated. It corresponds to the difference between 2022 and the construction year for buildings without energetic restoration. For the first energetic restoration per building, it corresponds to the difference between the year of energetic restoration and the construction year. For all other energetic restorations per building, it is equal to the difference between the years of two consecutive energetic restorations. Only lifetimes smaller than 152 years (corresponds to $Q_3 + (1.5 * IQR)$) were kept.
Finally, the categorical variables were encoded as a one-hot numeric array, and the numerical variables were standardized.

Another important data preparation step is the feature selection. Alsahaf et al. (2022) mentions that the presence of irrelevant and redundant features can degrade the performance of predictive models. If independent variables highly correlate in linear models, the estimates of regression coefficients are unreliable and unstable (Allison, 2012). To counteract these problems, I used a gradient boosting algorithm to calculate the feature importance of each independent variable and a correlation matrix. I selected the features with the highest importance score, which do not correlate. Finally, the choice was checked by calculating the variance inflation factor (VIF). According to Allison (2012) this factor is the most widely-used diagnostic for multicollinearity.

Modeling
Even though the restoration pressure is time-dependent, it is a good starting point to calculate a time-independent restoration pressure. In the second step, the time dependency was considered when thinking and researching suitable algorithms.

Time-independent restoration pressure:
A simple solution to calculate a time-independent restoration pressure is logistic regression. Logistic regression has been the most commonly used model for binary regression since about 1970 (J.S. Cramer, 2002). Although it is a classification method, it does not model the binary response variable but the probability that the response variable belongs to a particular class (James et al., 2013). And this probability could be suitable as a time-independent restoration pressure. In other words, buildings incorrectly classified as "energetically restored" could be said to have a high energetic restoration pressure.
Logistic regression is based on linear regression, where the logistic function is used to constrain the predicted values between 0 and 1. The logistic function p of a linear function f is given by

$$p(f(X)) = \frac{e^{f(x)}}{1 + e^{f(x)}} \tag{1}$$

The coefficients of the linear function ($\beta = (\beta_1, \cdots, \beta_n)$) are estimated with the training data and by minimizing the negative log-likelihood with regularization term $r(\beta)$:

$$\min_{\beta} C \sum_{i=1}^{n} (-y_i \log(p(f(X_i))) - (1 - y_i) log(1 - p(f(X_i)))) + r(\beta), \tag{2}$$

where y_i are the true classes for observation i and $p(f(X_i))$ is the prediction of the probability of the positive class $P(y_i = 1|X_i)$. The regularization term improves numerical stability. (James et al., 2013; Pedregosa et al., 2011a)

On the one hand, a baseline model was trained with the four most important features identified in the data preparation, and a model with all selected features (see Table 6). Shi et al. (2022) has shown that with imbalanced data, the minority class could be better identified by using resampling methods. But, it has the limitation of the high misclassification of negative cases. When modeling and evaluating energetic restoration pressure with logistic regression, it is

important to identify positive classes, with less weight given to the misclassification of negative cases. Moreover, only in 8.6% of the buildings an energetic restoration was identified. Therefore, I decided to train also models using resampling methods. There are mainly three types of resampling methods: oversampling methods, undersampling methods, and a combination of over- and undersampling methods. Undersampling is a technique to balance uneven datasets by deleting samples from the majority class. With oversampling, on the other hand, samples from the minority class are duplicated (Shi et al., 2022). Since a large amount of data is available in this study, I only tested different undersampling methods. More precisely, I used under-sampling with class weights and random under-sampling. Under-sampling with class weights assigns a weight to each class, in which the minority class gets a higher weight. While modeling, the class weight is used when calculating the negative log-likelihood. Therefore the model gets penalized more when it misclassifies a minority example than when it misclassifies a majority example (Agarwal, 2022). The amount of penalization is given by the weight, which can be calculated by the distribution of the target variable. In Random under-sampling, as the name indicates, a fraction of the majority class is randomly selected so that the data is balanced afterward.

For each model, the data was divided into train and test data. Then the best hyperparameters, like regularization term and algorithm used in the optimization problem, were selected by grid search, based on which the model was calculated on the training data. The test data was finally used to evaluate the model.

Time-dependent restoration pressure:
To calculate a time-dependent restoration pressure, I rely on survival analysis (Kleinbaum and Klein, 2012), which addresses the question, "how long would it be before a particular event occurs." This approach originated from population studies to model mortality probability is also called "time to event" analysis. Survival Analysis has been applied, besides the traditional death event, to various event types in different business domains, such as predictive maintenance in mechanical operations, customer retention, or cohort analysis (Harrison and Ansell, 2002; Hrnjica and Softic, 2021; Irigoyen et al., 2019). Volland et al. (2020) even applied survival analysis in the context of building renovations. They modeled the renovation probability of four energy-relevant building elements - heating systems, windows, facades, and roofs.
In the simplest models in survival analysis, the estimated survival function $S(t)$ depends only on time t and is defined as the probability that an event has not occurred by the time t. From the survival function, the hazard function can be derived, and vice versa, which is a measure of the risk of dying in an infinitesimally small time interval, given that the subject has survived up till the start of the interval. Extended models include the impact of predictor variables, called covariates, on the survival function (Pandey, 2020). According to McGregor et al. (2020), the CPH model is probably the most popular approach combining the covariates with the survival function.[4] It attempts to represent the hazard rate $h(t|x)$ as the product of the baseline hazard and the partial hazard:

$$h(t|x) = h_0(t)e^{\sum_{i=1}^{n}\beta_i x_i}. \tag{3}$$

The baseline hazard is the hazard when all variables x_i are nil. And when the value for a variable x_i changes, the partial hazard represents the change in the hazard ("Databricks", 2021). The formula of the hazard rate (3) implies the hazard assumption: the effect of the covariates on the hazard function is independent of time. Or in other words, the hazard ratio between two groups is proportional over time.

The hazard rate is a suitable measure to represent the time-dependent restoration pressure. It can be interpreted as the risk of an energetic restoration in an infinitesimally small time interval, given that the building has not been energetically restored up till the start of the interval. Since for some buildings, I have identified several energetic restorations, I assume that the different energetic restorations are independent and that a building is identified as a new building after

[4]Cox (1972) introduced a large family of models for estimating the hazard function. The CPH model is the simplest member of this family (Rodríguez, 2007).

an energetic restoration.[5] To estimate the hazard rate, I have used the CPH model from the python module scikit-survival.[6] For performance reasons, I calculated several hazard functions on a fraction of the data. To analyze the stability of the model and to avoid over-fitting, different random samples were chosen, which were further split into train and test data. Table 6 shows the predictor variables used in the CPH model and in the four different logistic regression models: baseline model (Baseline LR), full model (Full LR), under-sampling with class-weights model (CW u-LR) and random under-sampling model (Random u-LR). The renovation rate of the neighborhood was not used in the CPH model because it is time-dependent, and the construction year was replaced by the survival analysis required lifetime.

TABLE 6. Predictor Variables per Restoration Pressure Model

Feature	*Baseline LR*	*Full LR*	*CW u-LR*	*Random u-LR*	*CPH*
previous_renovation	x	x	x	x	x
energy_reference_area_pred	x	x	x	x	x
baujahr_upper	x	x	x	x	
canton	x	x	x	x	x
anz_wohnungen_egid		x	x	x	x
geb_klasse		x	x	x	x
anz_firmen_egid		x	x	x	x
geb_kategorie_gwr		x	x	x	x
bereits_getätigte_baukosten/gebäudevolumen		x	x	x	x
renovrate_neigh		x	x	x	

Evaluation

For both approaches (logistic regression and survival analysis), there are many widely used evaluation metrics but no exact concordance. However, to compare the different logistic regression models, I use some of the most frequently used evaluation metrics for binary classification: confusion matrix, F-Score, receiver operating characteristic (ROC) curve, and area under the (ROC) curve (AUC).

The confusion matrix forms the basis for other important measures like accuracy, recall, and precision. Accuracy is calculated by dividing all correct predicted classes through all predictions. Recall, also called true positive rate (TPR), measures the proportion of actual positives that are correctly identified:

$$Recall = \frac{TP}{TP + FN}$$

And precision is a measure of exactness:

$$Precision = \frac{TP}{TP + FP}$$

The F_1-score is a combination of the recall and the precision. In particular, it corresponds to the harmonic mean of them. The more general F-score, F_β, is the weighted harmonic mean of precision and recall (Shi et al., 2022):

$$F_\beta = \frac{(1 + \beta^2) \cdot Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall}$$

[5]Energetic restorations that took place within two years were combined and considered one restoration.

[6]Two different python libraries for survival analysis (lifelines and scikit-survival) were tested, whereat they differ in available models and built-in functions. Since the training with the CPH model from lifelines did not even converge with small data sets, I decided to use the CPH model from scikit-survival. A disadvantage of this module is that no built-in function for testing the Cox proportional assumption is available. However, if the goal is survival prediction, there is no need to worry about the assumption (Davidson-Pilon, 2019).

The F_β-score favors precision when β is smaller than 1, and recall when β is higher than 1.

Another common tool used to evaluate binary classifiers is the ROC curve and its AUC. The ROC curve plots the TPR against the false positive rate (FPR). The FPR is defined as $FP/(FP+TN)$). There is a trade-off between those two rates: the higher the TPR, the more FP are produced by the classifier (Géron, 2017). Different classifiers can be compared with the AUC. A value equal to 1 corresponds to a perfect classifier, whereas a value equal to 0.5 indicates a completely random classifier (James et al., 2013).

Depending on how classes are distributed or whether FN, respectively FP are more important, some performance measures are more suited than others. However, each metric depends on the chosen threshold, which splits the predicted data into classes. Therefore, I selected different thresholds and calculated the mentioned metrics to compare the different logistic regression models.

In survival analysis, the concordance index, also called C-index, is one of the most frequently used evaluation metrics and is a generalization of the ROC (Harrell et al., 1982; Kvamme et al., 2019; Pedregosa et al., 2011c). It is defined as the ratio of correctly ordered (concordant) pairs to comparable pairs (Pölsterl, n.d.). In other words, it estimates the probability that the predicted survival times of two randomly selected individuals have the same ordering as their true survival times. A perfect survival model will have a C-index equal to 1, and a model with a C-index equal to 0.5, predicts the relationship between risk and survival randomly (Albanese, 2022). Uno et al. (2011) has shown that the C-index is too optimistic with an increasing amount of censoring and therefore has introduced the Uno's-C, a C-index based on inverse probability of censoring weights.
Another interesting evaluation metric, which is especially useful for comparing the CPH model with logistic regression, is the time-dependent cumulative/dynamic ROC, calculated from cumulative cases and dynamic controls at each time point. Cumulative cases are all individuals that experienced an event prior to or at time t ($t_i \leq t$), whereas dynamic controls are those with $t_i > t$ (Pedregosa et al., 2011b). With this metric, it can be determined how well a model can distinguish subjects who fail by a given time ($t_i \leq t$) from subjects who fail after this time ($t_i > t$).

To compare the logistic regression with the survival model, I calculated the mean of the AUC of the time-dependent cumulative/dynamic ROC of the CPH model and compared it with the ROC AUC of the logistic regression models. Furthermore, the C-index and Uno's-C for the logistic regression were calculated. Instead of the survival time, the prediction of the probability of the positive class $P(y_i = 1|X_i)$ is used. And therefore, the index estimates the probability that the probability of the positive class of two randomly selected buildings have the same ordering as their true survival times, that is, the age of the building, and the age of the building at the time of energetic restoration, respectively.

Deployment
As mentioned in Section 3.2.1, the deployment step is not part of this thesis.

4. Results

Chapter 4.1 provides descriptive statistics of the data used in this study. The results of the classifications of the building applications are shown in Chapter 4.2. The evaluations of the classifications and the energetic restoration models are provided in Chapter 4.3 and 4.4, respectively.

4.1. Descriptive Statistics.

In this chapter, I provide short summaries of the distribution of the data used in this thesis. First, building applications from Docu Media Schweiz GmbH are considered, and afterward, the heat energy consumption data and the building characteristics.

4.1.1. *Building Applications.*

Table 7 gives an overview of missing values of the most important attributes in the building applications from Docu Media Schweiz GmbH. In 36.87% of the samples, the value of the attribute *text_ details* is missing. The text data of the building application (*projectdescription* and *text_ details*) were merged in the data preparation. The resulting attribute *_ text* contained only 12'391 samples with missing data, corresponding to 0.82% of the data.
In the variable *ga_ code* and *ga_ text*, values are missing in more than 30% of the samples. The attribute *ga_ text* encodes the attribute *ga_ code*, and since the number of missing values in these two fields do not match, I observe a data inconsistency.

Table 7. Missing Values of Building Application Attributes

Attribute	*Missing values*	*%*
baid	0	0.0
egid	147303	9.7
projectdescription	17008	1.12
projectdate	498	0.03
baukosten	239843	15.79
projectlanguage	12161	0.8
category_code	13242	0.87
category_text	12789	0.84
category_type	12789	0.84
ga_code	514720	33.89
ga_text	514382	33.86
text_details	560001	36.87
_text	12391	0.82

As shown in Table 8, most building applications are written in German. When considering the number of words per building application, a different distribution between the languages can be observed (see Figure 3). Building applications written in Italian contain on average more words than those written in French and German. A common feature of the distribution of the number of words per language is the right skewness. That is, most building applications are described with few words, and only a few building applications are described in detail. The median number of words in the German building application is nine, and the upper quartile is 20. This was an important insight for the further use of the text data for classification, especially for the decision which (machine learning) algorithms were used.

Table 8. Distribution of the Building Application Language

Language	*Number of building application*	*%*
German	1097972	72.28
French	336468	22.15
Italian	72395	4.77

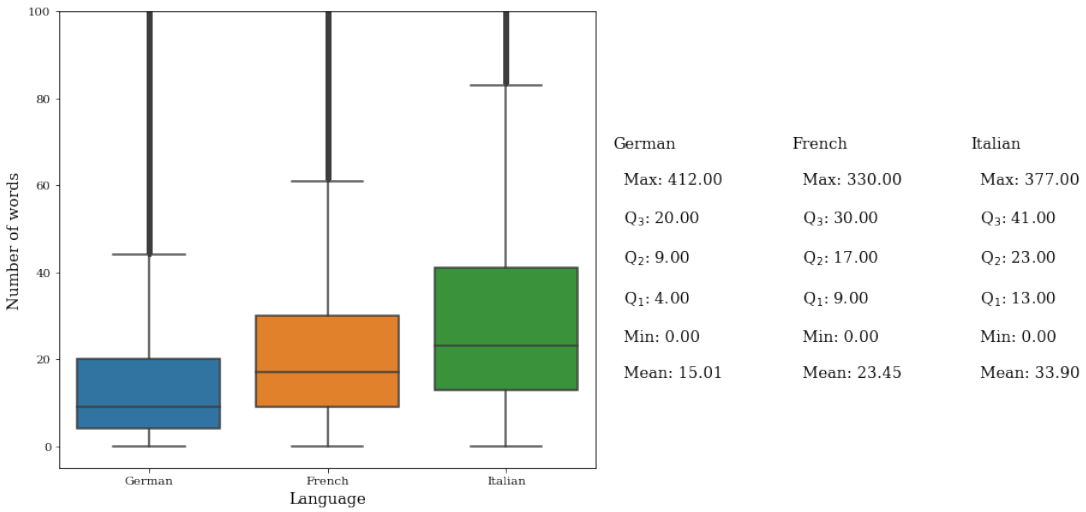


FIGURE 3. Distribution of Text Length per Language

Besides the text length, I am also interested in the content. I searched for structural measures belonging to energetic renovations (compare Table 1) in the German building applications description. Table 9 shows the top 20 structural measures based on their occurrence. The percentage is given in terms of the number of building applications written in German. By far, heat pumps (Wärmepumpe) occur the most frequently, followed by Sun protection devices (Sonnenschutz). Most energetic structural measures are used in less than 1% of the building applications. However, when searching for word parts like "dämmung" or " fassade, " the percentage increases to 3.07%, and 11.55%, respectively. It should be noted that words are now also counted that are not energetic structural measures, such as facade colors (Fassadenfarbe) or sound insulation (Schalldämmung).

TABLE 9. Most Common Energetic Renovation Activities in the Building Application Description

Structural measure	***Count***	***%***
Wärmepumpe	116406	10.60
Sonnenschutz	35061	3.19
Solaranlage	12729	1.156
Fernwärme	11946	1.09
Sonnenkollektoren	7647	0.70
Holzschnitzelheizung	2641	0.24
Photovoltaik	2043	0.19
Pelletheizung	1648	0.15
Holzheizung	1636	0.15
Kompaktfassade	1438	0.13
Wärmedämmung	1372	0.12
Fensterersatz	1173	0.11
Erdwärmesonde	814	0.07
Fassadenisolation	564	0.05
Dachisolation	508	0.05
Dachbegrünung	372	0.03
Fassadendämmung	318	0.03
Komfortlüftung	232	0.02
Holzfeuerung	180	0.02
Solarenergie	132	0.01
Dachdämmung	127	0.01

4.1.2. *Heat Energy Consumption.*

After removal of the outliers, that is, HEPI above 300 $kWh/m^2/year$ and blow 10 $kWh/m^2/year$, the distribution of the HEPI behaves weakly right-skewed, as can be seen in Figure 4. 50% of the remaining building have an HEPI between 78.2 $kWh/m^2/year$ and 124.5 $kWh/m^2/year$. The median is 101.70 $kWh/m^2/year$. For comparison, a renovated multi- and single-family house can have a maximum HEPI of 90 $kWh/m^2/year$ to apply for the minergie certification (Minergie Schweiz, 2022b).

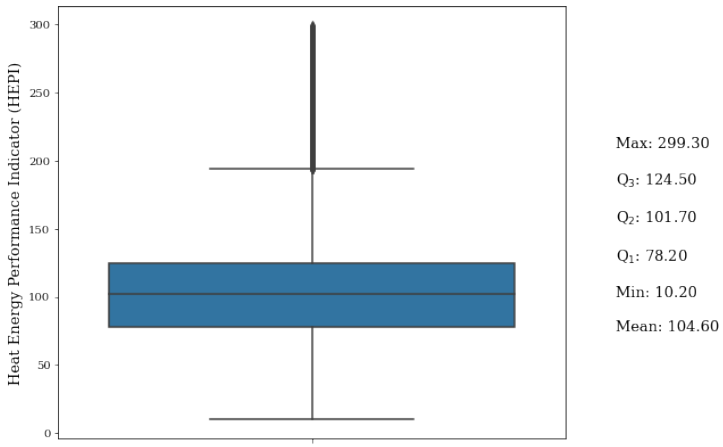


FIGURE 4. Distribution of the Heat Energy Performance Indicator (HEPI)

4.1.3. *Building Characteristics.*

An overview of the distribution and missing values of numeric building characteristics is given in Table 10. For 87.29% of existing buildings, the number of companies (anz_firmen_egid) is missing. Since the smallest recorded value is 1, it is assumed that missing values mean that there is no company in this building. With 70.89%, the attributes about renovations (renovationsjahr_lower and renovationsjahr_upper) also have a high proportion of missing values. Since this information was only recorded until December 2021 in the RBD, missing means that this building was not renovated before 2021 or that the renovation was not registered. Also noticeable is the largest value of *energy_reference_area_pred* compared to the largest value of *volume_egid*. Because the volume normally corresponds to three times the area, which is also approximately correct for the other key figures but not for the maximum.

TABLE 10. Distribution of Numerical Features of Existing Buildings

Attribute	***Missing values***	***%***	***mean***	***min***	Q_1	Q_2	Q_3	***max***
lat	0	0	47.01	45.82	46.69	47.12	47.40	47.80
long	0	0	8.04	5.97	7.39	8.06	8.74	10.49
baujahr_lower	195437	7.21	1948	1000	1926	1971	1992	2022
baujahr_upper	195437	7.21	1962	1000	1945	1971	1995	2022
renovationsjahr_lower	1920607	70.89	1994	1535	1986	1996	2007	2021
renovationsjahr_upper	1920607	70.89	1997	1535	1990	2000	2007	2021
anz_wohnungen_egid	0	0	1.75	0	0	1	1	460
anz_firmen_egid	2365073	87.29	1.9	1	1	1	2	345
anz_geschosse	758065	27.98	2.57	1	2	2	3	41
geb_flaeche2	0	0	174.3	0	56	103	176	99545
volume_egid	307987	11.37	1404	0	371	643	1254	1388096
energy_reference_area_pred	210087	7.75	446.3	0	166	222	410	248050
heating_degree_days_pred	0	0	3228.7	2126	2947	3082	3365	9449

Although Zurich (ZH) has the most inhabitants, it does not have the most buildings (Federal Statistical Office FSO, 2021). As Figure 5 shows, Bern (BE) has the most buildings. This can be explained by the fact that according to the building categories, 62% of the existing buildings are residential and 32% are non-residential (the remaining 6% are unclassified). The least buildings are in the cantons with the least inhabitants (Appenzell Inner Rhoden (AI), Nidwalden (NW), Obwalden (OW), and Uri (UR)).

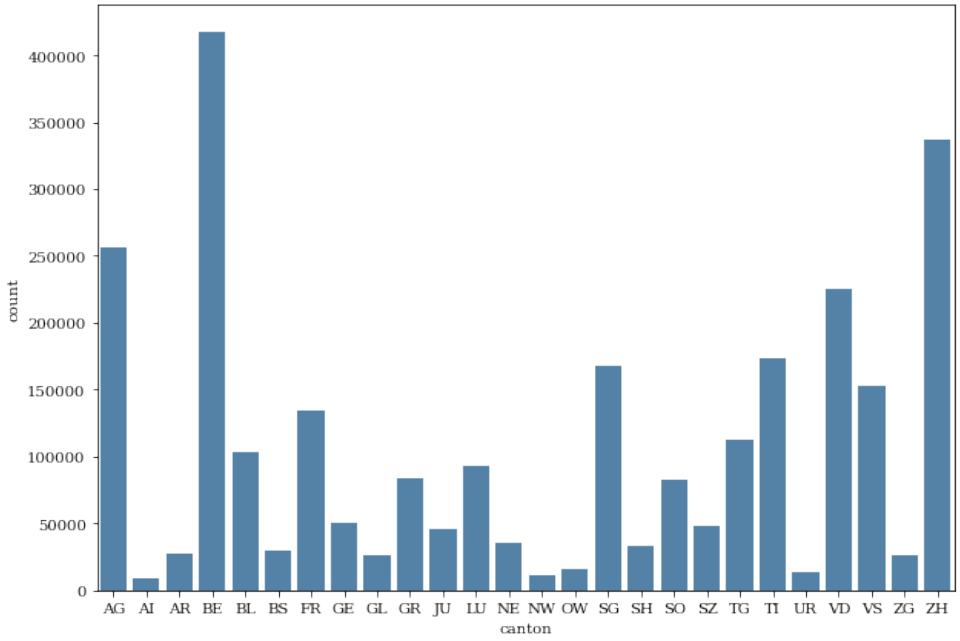


FIGURE 5. Number of Buildings per Canton

4.2. **Classification.**

In this chapter, I illustrate the results of the different classifications. As shown in Table 11, no building application was categorized as new construction and renovation. Furthermore, 2.23% of the building applications were identified as neither new construction, renovation, or demolition.

TABLE 11. Distribution of the Classified Application Type[7]

		New construction	
		True	***False***
Renovation	***True***	0%	56.93%
Renovation	***False***	40.84%	2.23%

Table 12 shows the distribution of the structural measure groups (building envelope and building technology) of renovations classified with the preexisting categories. Almost half of all building applications identified as renovations involve renovation activities affecting the building envelope. A renovation of the building technology takes place in almost 35% of the renovations. And in around 20%, the building envelope and the building services are renovated.

The distribution of the structural measures classified with the preexisting categories and their aggregations is given in Figure 6. 34% of the renovations were classified as energetic restoration. More than 30% of the renovations concern the roof, the window, or the insulation.

[7]This evaluation was created after the building applications identified as demolition or building applications with only irrelevant categories were deleted.

TABLE 12. Distribution of the Energetic Structural Measure Groups of Renovations

Structural measure	*count*	*%*
Building envelope	341990	49.68
Building technology	235461	34.21
Building envelope and technology	135247	19.64

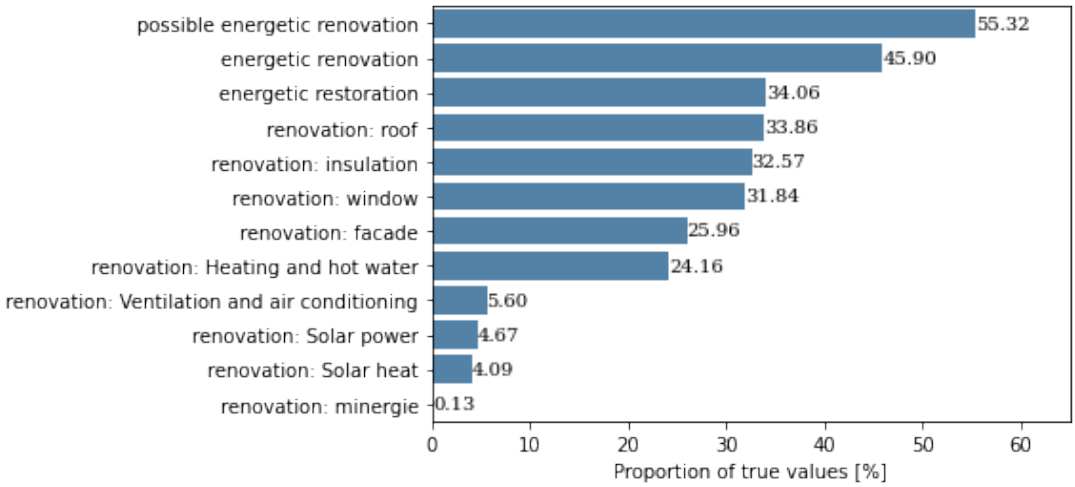


FIGURE 6. Distribution of the Classification with the Attribute ga_code

Using the building application descriptions, the defined structural measures were identified in fewer building applications than in the other classification approach. As shown in Figure 7, only 10.4% of the renovations were classified as energetic restoration. Heating and hot water renovations were identified only half as often when classifying with the text than with preexisting categories. Only in 10% of the renovations the building measure insulation was classified, corresponding to a quarter in comparison. In both classification approaches, less than 1% of the renovations were identified as Minergie certified. In section 4.3, I compare the two classification approaches in more detail using specific structural measures.

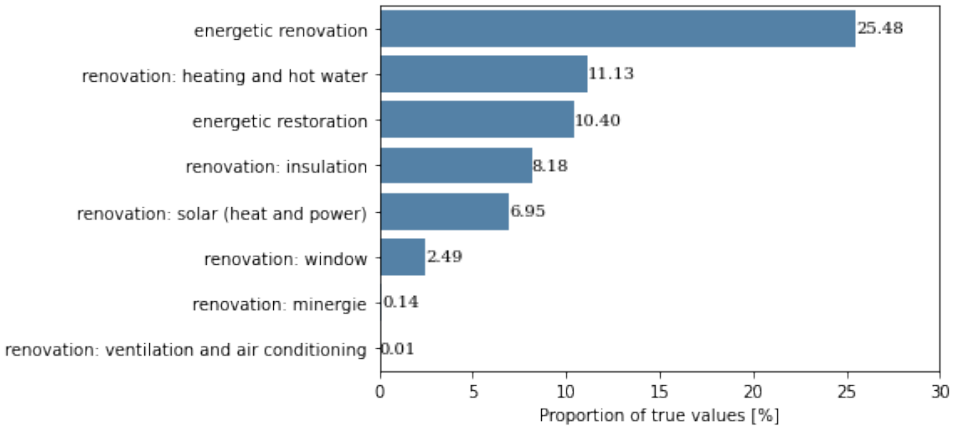


FIGURE 7. Distribution of the Classification Based on the Text Data

4.3. **Evaluations of Classification.**

I evaluated the classification based on preexisting categories and categories identified as possible energetic with the heat demand model of geoimpact AG. To evaluate the text classification, it is compared with the classification based on preexisting categories.

4.3.1. *Evaluation of the Classification Based on Preexisting Categories.*

To evaluate the energetic restoration class resulting from the classification based on preexisting categories, the energy-saving potential was predicted. For this purpose, the heat demand model from geoimpact AG, the HEC measures of individual buildings in the cities Geneva, St. Gallen, and Biel, and building characteristics were used. With the trained model and the test data, the HEC was predicted for each building, once for each class (energetic restored and non-energetic restored), and by leaving the other attribute values unchanged. From this, the HEPI was derived, and the difference was calculated. Analogously, the energy-saving potential of categories identified as possible energetic was calculated. In addition, the energy-saving potential of the minergie certification was predicted to have a reference value. Since buildings that are certified as minergie must have a low HEPI (Minergie Schweiz, 2022a). The partial dependency plots of all independent variables were created to get a complete overview of the model performance.

As expected, the partial dependency plot of the construction year showed a decreasing behavior for increasing construction years. This means younger buildings have on average a lower HEPI than older ones. Also, the attribute number of apartments shows an expected behavior in the partial dependency plot. The more apartments, the larger the predicted HEPI. This is an indication that the model works more or less well. In Appendix E.1, the resulting partial dependency plots of the model to evaluate the energetic restoration class are provided.

The results of the classification evaluation with the heat energy demand models are given in Table 13. Each row corresponds to the evaluation of a specific class. Once the model was trained with a linear SVR and once with a GBR. The HEPI difference, the MDAPE and MAPE were calculated for each model. The results of the reference class *minergie* are shown in the first row. As expected, the HEPI difference of this class *minergie* is negative for both approaches. According to the evaluation metrics (MDAPE and MAPE), the GBR performs slightly better than the linear SVR. It is striking that for classes with very low percentages of true values, the HEPI difference of the linear SVR model is strongly negative, whereas, in the GBR model, it is zero. In addition, some independent variables of the baseline model of geoimpact AG are linearly correlated, which may influence the model stability of linear models, and may cause over-fitting (Sundus et al., 2022, Wu, 2021). Therefore, I mainly refer to the results of the GBR model to discuss the influence of the classes on the HEPI. According to the heat energy demand model, an energetic restoration can save 5 $kWh/m^2/year$ on average. Furthermore, the following renovation activities on the building envelope show a negative HEPI difference: *Dachumbau, Dachausbau; Fassade ohne Detailangabe; Holz (Fassade); Mauerwerk verputzt; Fenster, Fenstertüren ohne Detailangaben; Sonnen-/Wetterschutz.* Somewhat surprisingly, *Sonn-/Wetterschutz* (sun/weather protection) and *Mauerwerk verputzt* (masonry plastered) appear in the list. The latter could be explained by the fact that this renovation activity is often combined with other activities on the building envelope, which leads to better insulation. In the case of *Sonn-/Wetterschutz*, it could be explained by the same argument, i.e., that this structural measure is combined with window replacement, which reduces energy consumption. Also exciting to see is that the class *Solarheizsystem* also achieves a negative HEPI difference. This can be explained by the fact that in the data, only the oil and gas consumption is contained, which decreases if a solar heating system is added.

TABLE 13. Prediction and Evaluation of Heat Energy Demand Model

Class	*Encoding*	*True values*	*%*	*HEPI Difference*		*MDAPE*		*MAPE*	
				LinSVR	*GBR*	*LinSVR*	*GBR*	*LinSVR*	*GBR*
minergie		230	1.13	-15.48	-12.52	20.74	18.66	34.48	31.29
energetic_restoration		2719	13.4	-4.87	-4.75	20.10	18.80	32.73	31.18
renovation_100	Dächer ohne Detailangaben	508	2.5	-1.35	0	21.35	18.78	35.81	31.26
renovation_101	Flachdach	403	1.99	-6.3	0	20.14	18.87	33.2	31.18
renovation_102	Schrägdach	272	1.34	2.63	0	20.37	18.86	31.4	31.18
renovation_104	Dachumbau, Dachausbau	408	2.01	-6.31	-2.85	21.07	19.02	35.04	31.12
renovation_151	Ziegel	214	1.05	-1.11	0	20.45	18.86	31.21	31.19
renovation_152	Faserzement	6	0.03	-24.86	0	20.39	18.86	33.69	31.17
renovation_200	Fassaden ohne Detailangaben	928	4.57	-4.62	-2.39	19.83	18.7	32.74	31.11
renovation_201	Metall, Stahl, Leichtmetall	130	0.64	-5.44	0	20.34	18.83	33.46	31.16
renovation_202	Holz	105	0.52	-6.53	-0.91	20.1	18.82	32.05	31.18
renovation_203	Naturstein	73	0.36	-6.87	0	20.32	18.87	33.29	31.19
renovation_204	Glas	127	0.63	-9.47	0	20.5	18.82	34.05	31.15
renovation_205	Mauerwerk verputzt	366	1.8	-8.52	-4.81	20.03	18.96	32.58	31.24
renovation_206	Fassadenelemente: Beton, Leichtbeton, Kunststein	35	0.17	-0.31	0	20.3	18.86	33.49	31.18
renovation_208	Faserzementplatten	33	0.16	-5.38	0	20.37	18.82	31.34	31.15
renovation_209	Keramik	1	0	-90.97	0	20.63	18.83	33.92	31.16
renovation_210	Sichtmauerwerk	4	0.02	-30.92	0	20.56	18.86	31.09	31.16
renovation_212	Sichtbeton	19	0.09	6.99	0	20.64	18.86	34.56	31.18
renovation_300	Fenster, Fenstertüren ohne Detailangaben	1604	7.9	-5.08	-3.41	20.63	18.77	34.17	31.15
renovation_503	Fernwärme	14	0.07	-33.11	0	20.6	18.87	34.32	31.18
renovation_504	Wärmepumpen	71	0.35	-4.46	0	20.12	18.83	32.59	31.13
renovation_506	Solarheizsysteme	417	2.06	-8.64	-8.29	20.41	18.44	33.99	31.25
renovation_507	Holzheizung	3	0.01	-28.26	0	20.44	18.83	31.16	31.16
renovation_508	Cheminées, Cheminéeöfen	34	0.17	-2.78	0	19.97	18.86	31.99	31.19
renovation_509	Bodenheizung	127	0.63	-13.13	-0.99	20.68	18.85	34.47	31.14
renovation_510	Heizkörper: Radiatoren, Heizwände	179	0.88	-8.45	0	20.37	18.83	31.42	31.15
renovation_511	Geothermie, Erdwärmesonden/-kollektoren	16	0.08	7.67	0	20.12	18.86	32.08	31.19
renovation_512	Holzschnitzelheizung	2	0.01	-34.02	0	20.36	18.83	33.52	31.16
renovation_513	Pelletheizung	6	0.03	-7.23	0	20.12	18.86	32.24	31.16
renovation_514	Kontrollierte Raumbelüftung, Komfortlüftung	18	0.09	-11.36	0	20.78	18.86	34.94	31.16
renovation_901	Klima	183	0.9	-11.1	-9.07	20.04	18.86	32.07	31.15
renovation_903	Sonnen- / Wetterschutz	1800	8.87	-4.59	-4.08	20.6	18.71	34.37	31.2
renovation_908	Kälteanlagen	327	1.61	-0.88	-0.03	20.47	18.82	30.92	31.19
renovation_917	Lüftung	806	3.97	-2.17	-0.01	20.45	18.75	31.19	31.2
renovation_1004	Solarenergie	93	0.46	4.28	0	20.11	18.86	31.9	31.15

4.3.2. ***Evaluation of the Classification by Text.***

For the evaluation of the classification based on text data, only building applications written in German are included. Several energetic renovation activities and the main classification into non-/energetic restoration are compared. I illustrate here only the evaluation of the energetic restoration and two structural measures (heat pump and insulation). The others can be found in Appendix E.2.

First, I compare the occurrence of heat pumps. In both classification approaches, the class *heat pump* has been identified about equally often. 9.5% of the renovation contains the word *wärmepumpe* in the description, and 9.7% contain the preexisting category for heat pump. However, I do not know if the two classification match. But the confusion matrix in Figure 8 contains exactly this information. The rows represent the preexisting class *heat pump*. The first row corresponds to all building applications classified as renovation that do not contain this class and the second row corresponds to all renovations containing this class. And the same applies to the columns that represent the label from the text classification. Building applications classified as renovation that contain the word *wärmepumpe* in the text belong to the second column. Most renovations either contain both labels or no label (lower right and upper left square). This is also reflected in the high TPR and true negative rate (TNR), which are greater than 90%.

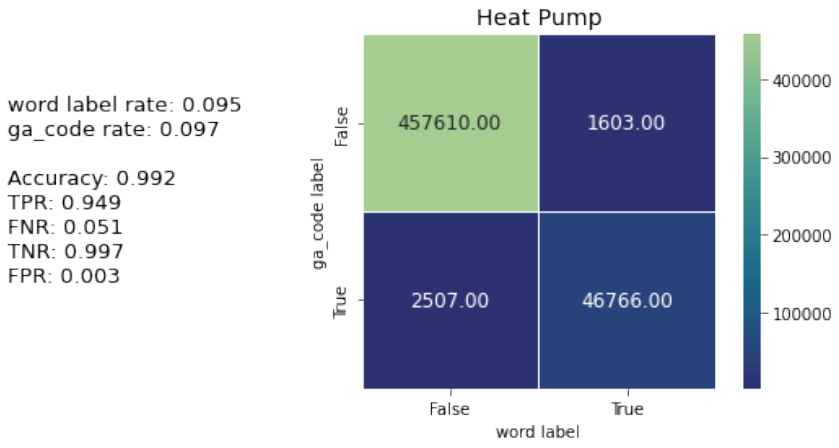


FIGURE 8. Comparison of Heat Pump Classification by Text and by ga_code

A somewhat different picture emerges in comparing the two classification approaches by the construction measure insulation. Only 8.2% of the building application classified as renovation contained a word of the word list obtained by the word parts *dämmung, dämung, dammung, dmmung, isolation, isolierung, fassade, dachbegrünung*. In contrast, 30% of the renovations include a class belonging to insulation (compare Table D.1 in Appendix D.1). As shown in Figure 9, most renovations do not contain the insulation in the text nor as a class. Therefore, the TNR is very high at 99.2%. In addition, 113'588 renovations were classified as insulation based on the preexisting categories but not by text, resulting in a false negative rate (FNR) of 74.5%.

Finally, I am interested in the classification of energetic restorations. In Figure 10, the renovation classified into energetic and non-energetic restoration with the text data (columns) is compared with the classification based on the preexisting categories (rows). Overall, 10% of the building applications classified as renovation are described by words identified as energetic restoration, and 30% contain a structural measure class identified as energetic restoration. In particular, based on the TPR and FNR, it can be concluded that more than 9% of all renovations have been identified as energetic restoration with both classification methods. It is also interesting to note that 1.1% of the renovations that were not classified as energetic restoration based on the preexisting classes were identified as energetic restoration via the text data.

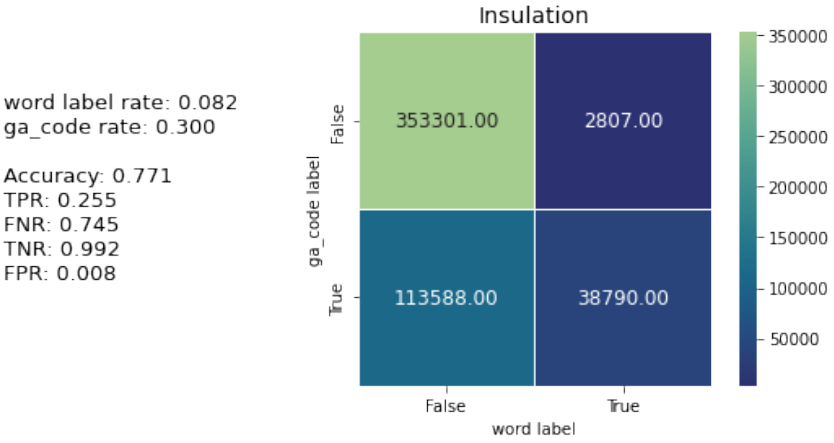


FIGURE 9. Comparison of Insulation Classification by Text and by ga_code

Since most renovations were not classified as energetic restoration, it is not surprising that the TN class in the confusion matrix is the largest (top left square).

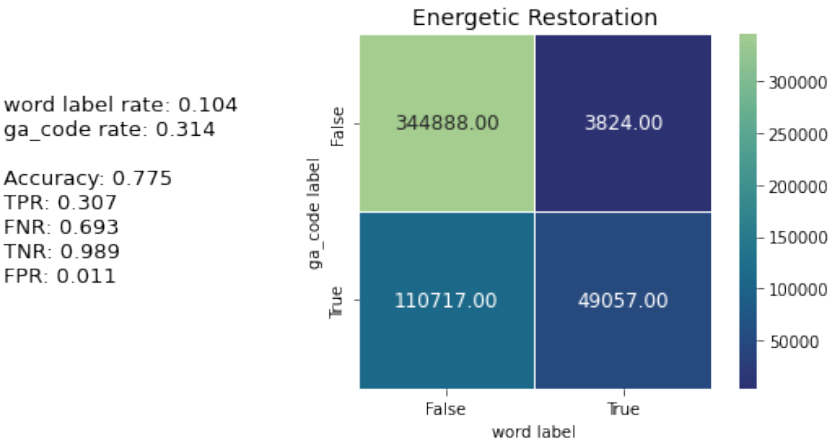


FIGURE 10. Comparison of Energetic Restoration Classification by Text and by ga_code

4.3.3. *Evaluation of the Final Classification.*

In this chapter, I discuss the evaluation of the final classification, which bases on the combination of the text classification approach and the classification based on the preexisting categories enriched with categories concerning the building envelope that showed an energy-saving potential in the heat energy demand model (i.e., classes with a negative HEPI difference in the GBR model, compare Chapter 4.3.1). Together with TEP Energy, 1'000 renovations identified and analyzed through surveys by TEP Energy were matched with existing building applications by the EGID. In particular, energetic restoration, window replacement, energetic facade renovation, and energetic roof renovation were compared, and the results are shown in Table 14. For more than 600 renovation activities, no building applications were available. The energetic roof renovation achieves the highest accuracy with 76% (regarding renovations with a building application). However, the TPR of energetic roof renovation is only 5%. But it should be mentioned that only two categories concerning the roof were identified as energetic and used for this classification. In all four compared renovation activities, the TNR is higher than the TPR. 50% of the energetic

restorations were correctly classified with the building applications.

TABLE 14. Comparison of Renovation Activities from Survey with Building Application Classifications[8]

		Building Application					
		Available		*Not Available*			
	Energetic Restoration	***True***	***False***		***Accuracy***	***TPR***	***TNR***
Survey	***True***	113	113	322	60%	50%	76%
Survey	***False***	34	108	311			
	Window Replacement						
Survey	***True***	69	103	265	62%	40%	82%
Survey	***False***	36	160	368			
	Energetic Facade Renov.						
Survey	***True***	55	57	79	73%	49%	83%
Survey	***False***	44	212	554			
	Energetic Roof Renov.						
Survey	***True***	4	80	87	76%	5%	96%
Survey	***False***	10	274	546			

4.4. **Evaluation of Restoration Pressure Models.**

To see the influence of the chosen thresholds on the evaluation metrics, I compare the metrics of the baseline logistic regression model based on different thresholds. Afterward, I compare the different logistic regressions on the selected evaluation metrics based on the particular best threshold of the F_3-score. Furthermore, I compare the logistic regression models with the CPH model by the C-index, Uno's C, and (mean-) ROC AUC. Finally, I show the feature importance of the best logistic regression model and the CPH model by considering the model coefficients.

4.4.1. *Threshold Comparison of Logistic Regression Baseline Model.*

I have compared five different thresholds: 0.5, 0.75 and the thresholds that achieved the best F_1-, F_2- and F_3-score. With these thresholds, the metrics accuracy, recall, precision, F_1-, F_2-, F_3-score and discrete ROC AUC of the logistic regression baseline model were calculated. As can be seen in Table 15, the accuracy cannot reflect the real performance of the given model. This is typical behavior of imbalanced data and is called the "accuracy paradox" (Valverde-Albacete and Peláez-Moreno, 2014). At the chosen threshold of 0.5 and 0.75, we get a very high accuracy of 91%, but if we look at the other metrics, we see that the models do not perform well. For example, they achieve only a ROC AUC of 0.5, which corresponds to a purely random classifier (Fawcett, 2006).

TABLE 15. Threshold Comparison of the Logistic Regression Baseline Model

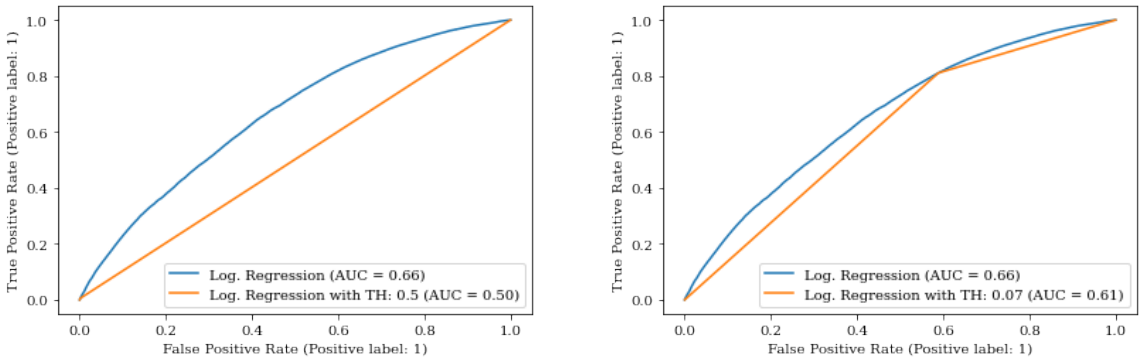
Model	***Threshold***	***Accuracy***	***Recall***	***Precision***	F_1	F_2	F_3	***discrete ROC AUC***[9]
0.75	default	0.91	0.002	0.28	0.004	0.002	0.002	0.5
0.5	default	0.91	0.005	0.26	0.01	0.006	0.005	0.5
0.09	F_1	0.7	0.48	0.14	0.21	0.32	0.39	0.6
0.07	F_2, F_3	0.45	0.81	0.11	0.2	0.36	0.5	0.61

[8]The metrics accuracy, TPR and TNR are calculated over the available building applications.
[9]The discrete ROC AUC is derived from the discrete ROC curve, and therefore belong to discrete predictions based on a given threshold.

As shown in Table 15, the accuracy and precision decrease with a decreasing threshold, while the other metrics increase. This is caused by the fact that with a smaller threshold the number of FP and TP classified samples increases, and the FN and TP classified samples decrease. This change in the number of FP and TP due to the choice of the threshold is also nicely shown in Figure 11. The blue curve, called the ROC curve, shows the FPR and TPR for all possible thresholds. The orange curve, which I will call discrete ROC curve, is based on three points:

- lower left point (0,0), which corresponds to the discrete model based on the threshold equal to 1 and therefore represents the strategy of never issuing a positive classification
- upper right point (1, 1), which corresponds to the discrete model based on the threshold equal to 0 and therefore represents the opposite strategy of unconditionally issuing positive classifications
- point defined by the FPR and TPR of the discrete model based on the chosen threshold (0.5, 0.007, respectively)

It is nice to see that the curve given by the threshold 0.07 (right Figure) is closer to the blue line than the curve given by the threshold 0.5. Therefore, the discrete model based on the threshold 0.07 achieves a better discrete ROC AUC than the other model based on the threshold 0.5.



(A) Threshold 0.5 (B) Threshold 0.07

FIGURE 11. The ROC Curves of Two Different Thresholds

Ultimately, we are interested in the FP classified samples (buildings that are wrongly classified as energetic restoration), and energetic restoration must be classified as such. The discrete model based on the threshold obtained by maximizing the F_2- and F_3-score achieved the best performance in this regard. Since I weigh recall more than precision, I decided to compare the different logistic regression models using the respective best thresholds of the F_3-score.

4.4.2. *Logistic Regression Evaluation and Comparison.*

I have tested four different logistic regression models:

- the baseline model: trained only on the four most important features,
- the full model: trained on all features,
- under-sampling with class-weights model: a model trained with all features and with class-weights to compensate for the imbalance of the two classes,
- a random under-sampling model: a model trained with all features but only on a sample of data that was balanced by random under-sampling.

Table 16 shows the evaluation of these four different logistic regression models based on the particular threshold that achieved the best F_3-score. It is interesting to see that all models achieve a similar F_3-score, but the corresponding thresholds are quite different. In particular, the thresholds of the models with under-sampling are significantly larger than the thresholds of the models trained on imbalanced data (baseline and full model). As a reminder, the threshold is used to assign the probabilities obtained from the model to the two classes. With a threshold

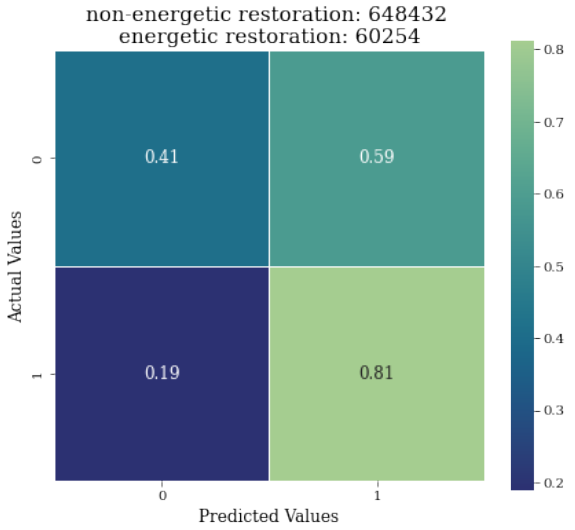
of 0.41, all buildings that reach a value greater than or equal to 0.41 are assigned to the energetic restoration class.

TABLE 16. Evaluation of Four Logistic Regression Models

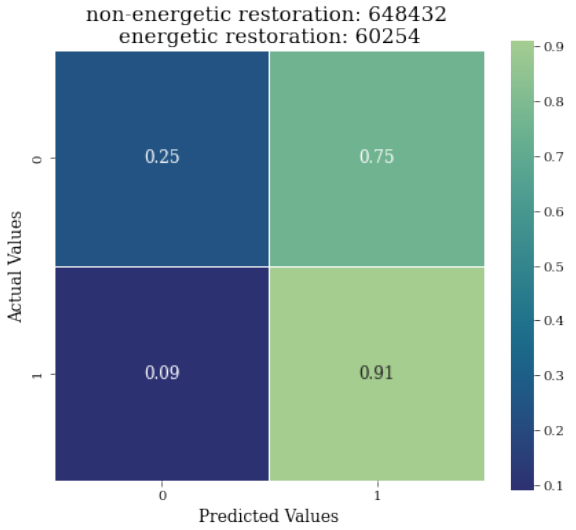
Model	***Threshold***	***Accuracy***	***Recall***	***Precision***	F_1	F_2	F_3	***discrete ROC AUC***
Baseline	0.07	0.45	0.81	0.11	0.20	0.36	0.50	0.61
Full	0.05	0.30	0.91	0.10	0.18	0.35	0.51	0.58
Under-sampling CW	0.41	0.45	0.84	0.12	0.20	0.37	0.52	0.62
Random Under-sampling	0.41	0.46	0.83	0.12	0.20	0.37	0.52	0.63

As we can see in Table 16, all models achieve rather a low precision, which is not surprising since the F_3-score considers the recall three times as important as precision. Both under-sampling models achieve very similar results. The under-sampling with class-weights achieves a slightly higher recall but a lower discrete ROC AUC and accuracy than the random under-sampling model. There is no model that stands out based on the selected thresholds and evaluation metrics. The biggest difference occurs in recall and accuracy. The full model achieves the highest recall, and the lowest accuracy. In addition, the full model based on the threshold 0.05 has the lowest discrete ROC AUC. However, when the ROC AUC of the models (which is independent of the threshold) is compared, the values are quite comparable: the under-sampling models yield the same and highest value with 0.7, followed by the full model with 0.69 and slightly behind the baseline model with a ROC AUC of 0.66.

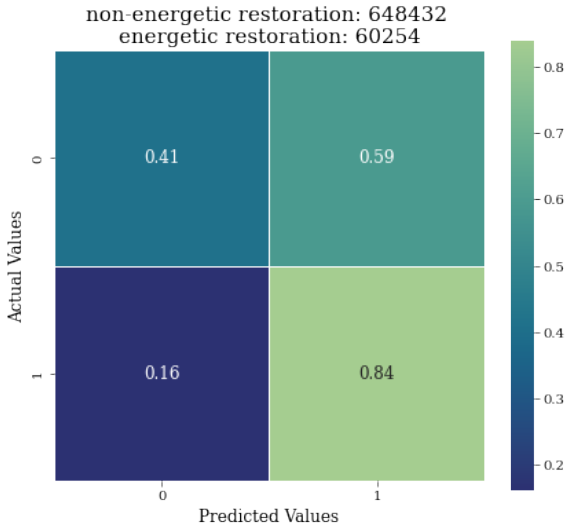
To get even more information about the model performance, we look at the confusion matrix of the four models in Figure 12. Per square, the numbers in the top left square represent the TNR, in the top right square the FPR, in the bottom left square the FNR, and in the bottom right square the TPR. In the heading, the number of test samples per class is given so that the absolute numbers can be calculated. Since the TPR is another name for recall, we again see that the full model properly classifies the most energetic restoration. However, the full model only correctly classifies a quarter of the non-energetic restorations. Thereby, with the full model, 75% of the non-energetic restored buildings would have a high energetic restoration pressure (more precise gradation of the pressure could be defined based on the model results, i.e., probabilities of the positive class). The other three models have a FPR between 57% and 59%, and thereby almost 60% of the non-energetic restored buildings would have a high restoration pressure according to these models.



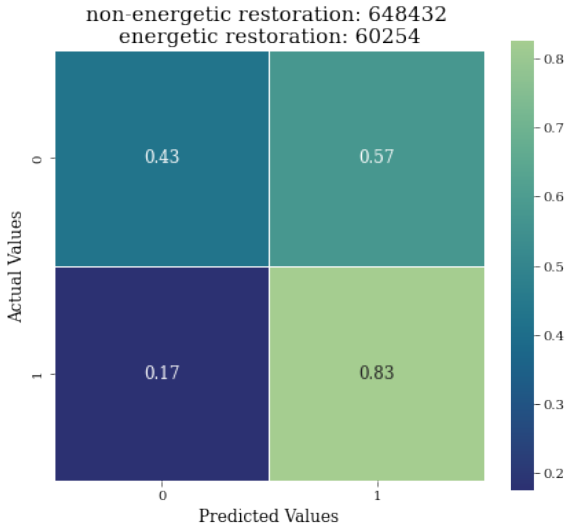
(A) Baseline Model (TH = 0.07)



(B) Full Model (TH = 0.05)



(C) Under-sampling CW Model (TH = 0.41)



(D) Random Under-sampling Model (TH = 0.41)

FIGURE 12. Confusion Matrices of Four Logistic Regression Models

4.4.3. *Cox Proportional Hazard and Logistic Regression Model Comparison.*

I evaluated the CPH model with the C-index, Uno's-C, and the mean of the time-dependent cumulative/dynamic ROC AUC. We compare the latter evaluation metric with the ROC AUC of the logistic regression models. The C-index and Uno's-C were adapted to logistic regression to have more comparable values. As shown in Table 17, the CPH model achieves the highest C-index. And since the amount of censoring data in the test data is high, the Uno's-C of the CPH model is a little smaller than the C-index, as expected. In the contrary, the Uno's-C of the logistic regression models is higher than the corresponding C-index, that are all close to 0.5. However, if we compare the (mean-) ROC AUC of the models, a somewhat different picture emerges. All models have a very similar ROC AUC, whereby the logistic regression models with under-sampling reach the highest value of 0.7.

TABLE 17. Evaluation of Cox Proportional Hazard and Logistic Regression Models

Model	*C-index*	*Uno's-C*	*(mean-) ROC AUC*
Cox proportional hazard model	0.70	0.68	0.69
Logistic regression baseline model	0.43	0.49	0.66
Logistic regression full model	0.56	0.61	0.69
Logistic regression with CW under-sampling	0.52	0.58	0.70
Logistic regression with random under-sampling	0.53	0.60	0.70

4.4.4. *Feature Importance.*

Another interesting result of a model is the feature importance, that is, the effect of a feature on predicting the output. In logistic regression and the CPH model, the coefficients of the feature in the model are responsible for their influence on predicting the energetic restoration pressure. Since I have not checked the proportional hazard assumption, and neither had the coefficient significance available in both methods, the results in this section must be taken with a grain of salt.[10] Nevertheless, I present here the coefficients with an absolute value $\geq |0.01|$ of one logistic regression model and the CPH model in order to check, if they behave as expected. Since the full model of the logistic regression model achieved the highest recall and in all other evaluation metrics it performed very similarly to the other ones, I decided to show the feature importance of this model. The other feature importance plots are shown in Appendix E.3. As can be seen in Figure 13, the construction year has a negative coefficient, which corresponds to the expectation that older buildings should have a higher pressure. Also as expected, the coefficient of the renovation rate in the neighborhood is positive, which means that the higher the rate of renovations in the neighborhood, the greater the restoration pressure. Furthermore, renovation that took place before 2004 decreases the probability of later energetic restorations.

Figure 14 shows the coefficients of the CPH model. All classes of the building category have the largest positive effect. The building classes for restaurants and bars in buildings without residential use, museums, and libraries have the largest negative effect. The hazard of an energetic restoration decreases with increasing costs per building volume of previous non-energetic renovations, or if a renovation took place before 2004.

[10] If the goal is prediction, then the proportional hazard assumption does not need to be tested. However, there are legitimate reasons to assume that all datasets will violate the proportional hazards assumption (Davidson-Pilon, 2019). Furthermore, I used the python packages sklearn and sksurv, which do not provide coefficient significance. Of course, this calculation could also be developed on its own. However, due to time constraints, this was omitted.

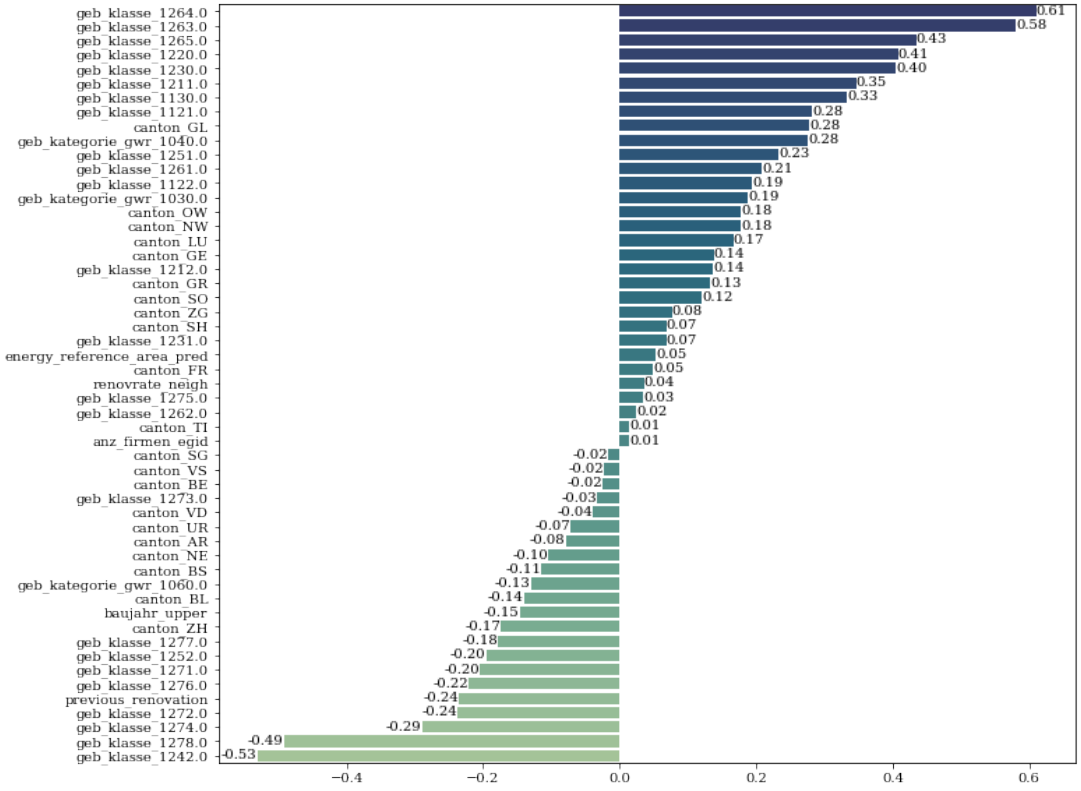


FIGURE 13. Feature Importance of Logistic Regression Full Model

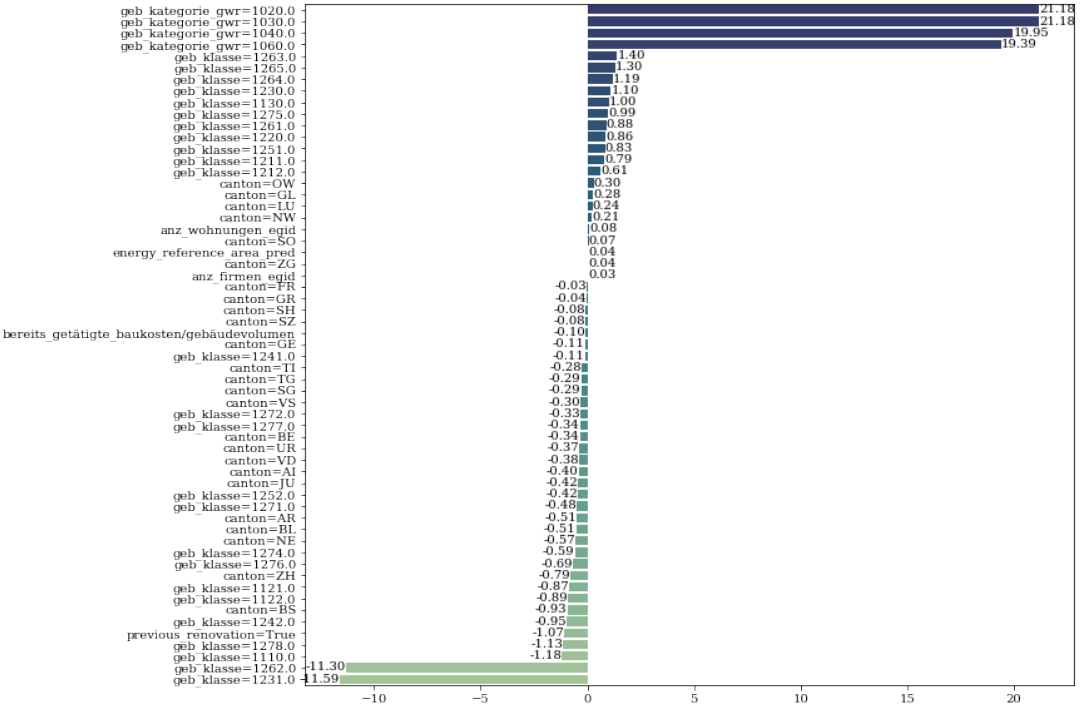


FIGURE 14. Feature Importance of Cox Proportional Hazard Model

5. Discussion and Conclusion

In this chapter, I summarize the challenges and the main findings. In doing so, I chronologically follow the thesis's structure and answer the research questions. Finally, I draw a conclusion and make some recommendations.

5.1. Discussion of Building Application Classifications.

The building applications have raised some challenges. Identifying the application type (renovation and new construction) was not unambiguous and is a possible source of error. However, I have developed a rule which assigned the class renovation rather conservatively so that no ambiguity arises.
Other challenges appeared with the free text data of the building applications. Since the different languages, the poor quality related to spelling and the length of the text has caused some restrictions. Due to the many spelling errors and domain-specific language, some text preparation steps, like lemmatization and spelling correction, have worked poorly and were useless. In particular, a largely manual effort would be necessary to correct the spelling errors. Furthermore, the text length of the single building application is rather short, as shown in Figure 3. For example, the median number of words of building applications written in German is nine. Since machine learning methods usually fail to achieve the desired accuracy when classifying short text data (McCartney et al., 2017; Phan et al., 2008), I faced another challenge. Consequently, rule-based systems based on literature about energetic renovation remained as appropriate classification approaches. Building application descriptions and preexisting categories contained information on structural measures and were therefore used to develop rule-based systems.

According to the heat energy demand model, the energetic restoration identified by the developed rule-based system based on preexisting categories showed on average, an energy saving potential of almost 5 $kWh/m^2/year$, compare Table 13. Since 75% of the buildings that were used to train the heat demand model have an HEPI lower than 124 $kWh/m^2/year$, 5 $kWh/m^2/year$ seems to be quite low. Streicher et al. (2019) states that only with deep building envelope renovations the final energy demand of the Swiss building stock can be reduced by 57%. Since the mean HEPI of the available energy consumption data equates to 104 $kWh/m^2/year$, on average, the reduction of the energy demand could be lowered by almost 60 $kWh/m^2/year$, if according to Streicher et al. (2019) a deep building envelope renovation would have taken place. The big difference between the average saving potential between our estimations and from the literature can have several reasons. For example, I have identified a building application as energetic restoration if at least one structural measure class identified as energetic was included. Therefore the potential for energy consumption reduction is probably lower compared to a deep building envelope renovation. Furthermore, the information content of the preexisting categories of the building applications exhibits gaps. Although I only included structural measures with an energy-saving potential according to technical literature, the degree of the structural measurement is not represented in the category. Depending on the basic structure of the building and insulation material, insulation thicknesses of 14 to 25 cm are necessary to make a house sufficiently energy-efficient (Swiss Federal Office of Energy SFOE, 2022). Also, the energy efficiency of windows can differ, and the number of windows replaced in a building influences the degree of the energy-saving potential. Jakob et al. (2014) found out through a survey that often not all windows are replaced, but only parts of them.
Not only the content of the identified energetic restoration can influence the result, but also the wrong classifications as well as missing building applications. For more than 50% of energetic restorations, no building application was found, as the random sample comparison of the building applications with the surveys on renovation activities of TEP Energy has shown (compare Chapter 4.3.3). This further increases the rate of unidentified energetic restorations, which can further worsen the classification evaluation with the heat energy demand model.
However, the reference class *minergie*, shows a lower estimated energy-saving potential, as expected. With an estimated energy consumption saving potential of 12.5 $kWh/m^2/year$,

Minergie-certified buildings are only a little better than energetic restoration and still far from the energy demand reduction potential of deep envelope renovation derived from the literature (Streicher et al., 2019). One possible explanation is based on the heat energy consumption data used in the heat energy demand model. The upper quartile of the HEPI of the building used to train and test the model is 124 $kWh/m^2/year$. And, depending on the building category, the energy consumption of renovated buildings must not exceed 90 to 125 $kWh/m^2/year$ to be able to receive a Minergie certificate according to Minergie Schweiz (2022b). I conclude that a large proportion of the buildings used in the heat energy demand model were already "good" buildings in terms of energy consumption. And I think when an already "good" building gets energetically restored, the potential for energy consumption reduction is not as high as for "bad" buildings. And therefore, the heat demand model tends to underestimate the influence of specific renovation activities.

Considering that the heat energy demand model underestimates the energy-saving potential (either due to the tendency of too many too "good" buildings in the heat energy consumption data or due to the classification of the building application itself), I decided to add further structural measures that showed an energy consumption saving potential in the heat energy demand model to the set of rules for the classification of the energetic restorations based on the preexisting categories.

The rule-based system with the text data was only defined based on building applications written in German, which concerns almost 73% of the buildings. The evaluation of this classification consists of a comparison with the result of the rule-based system based on the preexisting categories. Since the heat energy consumption data used in the heat energy demand model by the majority concerns buildings in the French-speaking part of Switzerland, this evaluation approach was not appropriate. Depending on the compared class, the two classifications have better or slightly less well coincided (compare Figure 8 and 9). As we have seen in Figure 10, most building applications that were classified as non-energetic restoration with the text data were also classified as non-energetic restoration by the classification method based on the preexisting categories. However, the TPR was quite low at 30.7%. When considering the building applications that have fallen into the FP class, no category is available in more than half of the building applications. And also when considering examples of the FN class, nothing conspicuous can be recognized, except that the text contains no information on energetic construction measures of the building envelope. Consequently, the two approaches seem to complement each other well, and a combination seems to be a plausible approach. Therefore, a combination of the two rule-based systems was applied in the final classification, which formed the basis for calculating the energetic restoration pressure.

The evaluation of the final classification based on survey data from TEP Energy was slightly sobering. On the one hand, because of the 1'000 renovations identified by the survey, there is a building application only in less than 400 cases. And on the other hand, the correspondence of the renovation activities in the rest has reached only a maximum accuracy of 76% (compare Table 4.3.3). The fact that in less than 60% of the compared renovations, there is no building application can have several reasons. First, the question arises whether a building permit is necessary for the renovation activity. If required, it can still be that no building permit was requested, or Docu Media Schweiz GmbH has not recorded the building application. Given that the building application is available, errors can still occur when linking addresses and assigning the EGID. The rather bad accuracy of the building application classification can also have different causes. Either the classification is bad, and the rule-based systems need to be adjusted, or the building application does not contain the corresponding information. The latter could be because not all renovation activities require a permit, and maybe therefore some renovation activities are not added to a building application. However, the low TPR of the energetic roof and facade renovation can be explained by the fact that the building application classification did not distinguish between the insulation of individual building elements. Therefore, this information was not used to classify energetic roof and facade renovations.

5.2. **Discussion of the Restoration Pressure Modeling.**

The question "How can energetic restoration be predicted as a time-dependent process?" is essential but also challenging. Therefore, I have first estimated a time-independent restoration pressure as a restoration probability with the logistic regression method. Such simple approaches form a good baseline for evaluating more complex approaches. To include the time-dependence of the dependent variable, I have used a survival analysis method, the CPH model. As stated in chapter 3.2.2, this "time to event" analysis is used in various fields where an event must be predicted as a function of time. Renovations, especially energetic restorations, can be identified as such events, as Volland et al. (2020) has already pointed out.

In addition to selecting a suitable model, the question of suitable evaluation metrics is central. As described in chapter 3.2.2, many commonly used evaluation metrics for classification problems exist. Since we are interested in the FP classified samples (buildings that are wrongly classified as energetic restoration), and it is important that energetic restorations are classified as such, the recall and die ROC AUC are appropriate to assess the performance of the logistic regression models.
To evaluate the CPH model, I identified Uno's-C as the most appropriate evaluation metric. With the increasing amount of censoring, the Uno's-C is more accurate than the C-index (Uno et al., 2011), and therefore in our case more appropriate. To be able to make a comparison with the logistic regression models, the time-dependent cumulative/dynamic ROC AUC was calculated for the CPH model. Furthermore, I calculated the C-index and Uno's-C of the predictions of the logistic regression models. The results (see Table 17) show that these adaptations did not work well. Contrary to the expectation, the C-index value is higher than the Uno's-C value. Furthermore, the values are close to 0.5, which corresponds to a random prediction and thereby contradicts the ROC AUC, which states that predictions are better than random. From this, I conclude that the C-index and Uno's-C are inappropriate for evaluating the logistic regression.
So, in summary, the most appropriate metrics to measure the model's performance differ per model. For the survival analysis, I prefer the Uno's C, since it is easier to interpret than the time-dependent cumulative/dynamic ROC AUC and evaluates the model as a whole. And as already mentioned, to evaluate the logistic regression model in our use case, I prefer the recall and ROC AUC. However, to compare both approaches, the most suitable metric is the (mean-) ROC AUC.

To answer the main research question, "Which approach achieves the best performance in predicting energetic restoration of a building?" I considered the evaluation of the models, in particular the (mean-) ROC AUC in Table 17. The CPH model performed slightly worse than the two logistic regressions with under-sampling, which achieved the best ROC AUC of 0.70. On the other hand, the CPH model has the advantage that it models the energetic restoration pressure (given as hazard) in a time-dependent way. In contrast, only a time-independent energetic restoration pressure can be estimated with the logistic regression models. Since this advantage is even a requirement, and the difference in performance minimal, the CPH model is in my opinion better suited to model the energetic restoration pressure. But since the achieved Uno's-C of the CPH model is quite far from the perfect model having an Uno's-C of one, it would be exciting to find out if other models of survival analysis perform better. However, concerning the quality of the energetic restoration classification, the results should be taken with caution. In particular, the approaches used in this thesis and more complex models should be investigated on more reliable data. An example of a more complex model is the Cox's time-varying proportional hazard model, which can include time-dependent covariates. Or the parametric CPH, which allows more flexibility by making the baseline hazard parametric ("Databricks", 2021). Moreover, there are models available that can account for complex, non-linear relationships between survival and features, like Survival Support Vector Machine, Random Survival Forests, and Gradient Boosting for survival analysis (Pedregosa et al., 2011d). And even a deep learning approach to survival analysis called DeepHit introduced by Lee et al. (2018) is available.
A further possibility to improve the model would be a more differentiated consideration of the

energetic restorations. That is, not to calculate a general energetic restoration pressure, but the pressure for individual energetic restoration activities, such as facade insulation or window replacement. Because the renovation probabilities of individual energetic measures such as facade and windows differ according to Volland et al. (2020), it is maybe more appropriate to model their restoration pressure individually.
The last thing that should be mentioned is the assumptions made when the data were prepared for the CPH model. Several building applications classified as energetic restoration can belong to a building, and these different energetic restorations were considered independent. In particular, the building was considered a new building after each energetic restoration. This is very hazardous since a building application was identified as such if at least one energetic restoration activity was included. If, as already mentioned, a differentiated restoration pressure were calculated, this problem would be remedied. However, the assumption of independence could still cause problems. This problem could be handled by using only one energetic restoration or techniques to model recurrent events, like Prentice-Williams-Peterson models (Prentice et al., 1981).

5.3. **Conclusion.**

In conclusion, it can be said that building applications are not a viable alternative to surveys to obtain information about renovations, in particular, to identify energetic restorations. I would recommend studying for which renovation activities a building application is necessary. For renovation activities that require a permit, it would then have to be examined whether building applications contain the necessary information content. The rule-based approach presented in this thesis would be suitable for this purpose but would certainly need to be adapted and possibly extended.
Considering the poor data quality of the building application classification, the calculation of a restoration pressure based on these data is rather not recommended. Furthermore, I would advise training and re-evaluating the different approaches on more accurate data to obtain reliable results. Nevertheless, in my opinion, the survival analysis, in particular, the CPH model is a suitable approach to predict the pressure of renovation activities. However, I would recommend calculating the restoration pressure in a more differentiated way, that is, for specific energetic restoration activities. Furthermore, different survival analysis methods, like Cox's time-varying proportional hazard model and Survival Support Vector Machine, and recurrent events methods should be compared.

Acknowledgements

I would like to thank TEP Energy, in particular Martin Jakob and Jonas Müller for their contribution to the evaluation of the classification of building applications. The reconciliation of building applications with TEP Energy's survey data has made a significant contribution to assessing the potential of building applications as a source of information for renovations. I would also like to thank my supervisors Philipp Schütz and Esther Linder for all their help and advice with this thesis. Finally, I would like to express my gratitude to Thilo Weber and geoimpact AG for the good cooperation and support.

References

Agarwal, R. (2022, April 26). *The 5 most useful techniques to handle imbalanced datasets* [Medium]. Retrieved November 24, 2022, from https://towardsdatascience.com/the-5-most-useful-techniques-to-handle-imbalanced-datasets-6cdba096d55a

Albanese, N. C. (2022, June 26). *How to evaluate survival analysis models* [Medium]. Retrieved November 26, 2022, from https://towardsdatascience.com/how-to-evaluate-survival-analysis-models-dd67bc10caae

Allison, P. (2012, September 10). *When can you safely ignore multicollinearity?* [Statistical horizons]. Retrieved November 24, 2022, from https://statisticalhorizons.com/multicollinearity/

Alsahaf, A., Petkov, N., Shenoy, V., & Azzopardi, G. (2022). A framework for feature selection through boosting. *Expert Systems with Applications*, *187*, 115895. https://doi.org/10.1016/j.eswa.2021.115895

Bengfort, B., Bilbro, R., & Ojeda, T. (2018, June). *Applied text analysis with python*. O'Reilly Media, Inc.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Cross-industry standard process for data mining [Page Version ID: 1081205231]. (2022, April 5). In *Wikipedia*. Retrieved May 7, 2022, from https://en.wikipedia.org/w/index.php?title=Cross-industry_standard_process_for_data_mining&oldid=1081205231

Databricks. (2021). Retrieved December 2, 2022, from https://www.databricks.com/notebooks/telco-accel/03_cox_proportional_hazards.html

Davidson-Pilon, C. (2019). *Lifelines: Testing the proportional hazard assumptions* [Journal of open source software] [Publisher: The Open Journal]. Retrieved December 2, 2012, from https://lifelines.readthedocs.io/en/latest/jupyter_notebooks/Proportional%20hazard%20assumption.html#Do-I-need-to-care-about-the-proportional-hazard-assumption?

Düntsch, I., & Gediga, G. (2019). Confusion matrices and rough set data analysis. *Journal of Physics: Conference Series*, *1229*(1), 012055. https://doi.org/10.1088/1742-6596/1229/1/012055

Earth observatory [2021 continued earth's warming trend]. (2021). Retrieved April 2, 2022, from https://earthobservatory.nasa.gov/images/149321/2021-continued-earths-warming-trend

Ebert Stoll, B. (n.d.). *Die Gebäudehülle - der Königsweg zum Heizungsersatz* [Forumenergie]. https://forumenergie.ch/images/fez/anlaesse/fez/special/2020/pdf/02_FEZSepcial_Folien_EbertStoll_20200429.pdf

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Federal Laboratory for Materials Testing and Research. (2021). Energy renovation: First sort, then refurbish. Retrieved April 30, 2022, from https://www.admin.ch/gov/en/start/documentation/media-releases.msg-id-84812.html

Federal Office for the Environment FOEN. (2018, August 21). *Das Übereinkommen von Paris*. Retrieved April 2, 2022, from https://www.bafu.admin.ch/bafu/de/home/themen/thema-klima/klimawandel-stoppen-und-folgen-meistern/klima--internationales/das-uebereinkommen-von-paris.html

Federal Office for the Environment FOEN. (2020, December 17). *Gebäude*. Retrieved April 2, 2022, from https://www.bafu.admin.ch/bafu/de/home/themen/thema-klima/klimawandel-stoppen-und-folgen-meistern/schweizer-klimapolitik/gebaeude.html

Federal Office for the Environment FOEN. (2021, July 30). *Massnahmen, die mit dem Nein zum CO2-Gesetz per 1. Januar 2022 auslaufen oder beschränkt werden*. Retrieved April 2, 2022, from https://www.bafu.admin.ch/bafu/de/home/themen/thema-klima/klima--rechtliche-grundlagen/totalrevision-co2-gesetz/auslaufende-massnahmen.html

Federal Office for the Environment FOEN. (2022, November 4). *Indikator Klima*. Retrieved May 19, 2022, from https://www.bafu.admin.ch/bafu/de/home/themen/thema-klima/klima--daten--indikatoren-und-karten/klima--indikatoren/indikator-klima.html

Federal Statistical Office FSO. (2018, December 20). *GWR: Merkmalskatalog* [Federal Statistical Office FSO]. Retrieved December 13, 2022, from https://www.bfs.admin.ch/news/de/2018-0221

Federal Statistical Office FSO. (2021). *Ausgewählte indikatoren im regionalen vergleich, 2021 (kantone)*. Retrieved November 27, 2022, from https://www.bfs.admin.ch/bfs/de/home/statistiken/regionalstatistik/regionale-portraets-kennzahlen/kantone.html

Galimshina, A., Moustapha, M., Hollberg, A., Padey, P., Lasvaux, S., Sudret, B., & Habert, G. (2020). Statistical method to identify robust building renovation choices for environmental and economic performance. *Building and Environment*, *183*, 107143. https://doi.org/10.1016/j.buildenv.2020.107143

Gebäudehülle Schweiz. (2010). Der Königsweg der Gebäudesanierung. https://weberdach.ch/wp-content/uploads/2019/02/Broschuere_Gebauudesanierung_Koenigsweg_2010.pdf

geoimpact AG. (n.d.-a). *Info Gebäude* [Swiss Energy Planning SEP]. Retrieved April 9, 2022, from https://www.swissenergyplanning.ch/gebaeudeinfo-fr-ch

geoimpact AG. (n.d.-b). *Plattform* [Swiss Energy Planning SEP]. Retrieved May 7, 2022, from https://www.swissenergyplanning.ch/plattform

Géron, A. (2017, March). *Hands-on machine learning with scikit-learn, keras, and TensorFlow* (1st Edition). O'Reilly Media.

Grosan, C., & Abraham, A. (2011, July 29). *Intelligent systems: A modern approach* [Google-Books-ID: c1fzgQj5lhkC]. Springer Science & Business Media.

Gubser, A. (2021, December 23). *Modelling the energy demand of individual buildings in switzerland* (Doctoral dissertation). Lucerne University of Applied Science and Art.

Hansen, J. (2012). Opinion | game over for the climate. *The New York Times*. Retrieved May 7, 2022, from https://www.nytimes.com/2012/05/10/opinion/game-over-for-the-climate.html

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA*, *247*(18), 2543–2546.

Harrison, T., & Ansell, J. (2002). Customer retention in the insurance industry: Using survival analysis to predict cross-selling opportunities. *Journal of Financial Services Marketing*, *6*(3), 229–239. https://doi.org/10.1057/palgrave.fsm.4770054

Hrnjica, B., & Softic, S. (2021). *The survival analysis for a PredictiveMaintenance in manufacturing*. Springer. Retrieved December 4, 2022, from https://doi.org/10.1007/978-3-030-85906-0_9

Irigoyen, M., Porras-Segovia, A., Galván, L., Puigdevall, M., Giner, L., De Leon, S., & Baca-García, E. (2019). Predictors of re-attempt in a cohort of suicide attempters: A survival analysis. *Journal of Affective Disorders*, *247*, 20–28. https://doi.org/10.1016/j.jad.2018.12.050

Jakob, M., Catenazzi, G., Sunarjo, B., Müller, J., & Weinberg, L. (2021, July). Kantonale energiekennzahlen und CO2-emissionen im gebäudebereich. Retrieved December 15, 2022, from https://www.tep-energy.ch/docs/de_en/p1111_TEP_Kantonale_EnergieKennzahlen-co2-Emissionen-Gebaudebereich.pdf

Jakob, M., Martius, G., Catenazzi, G., & Berleth, H. (2014, February). Energetische erneuerungsraten im gebäudebereich - synthesebericht zu gebäudehülle und heizanlagen. Retrieved November 15, 2022, from https://pubdb.bfe.admin.ch/de/publication/download/7387

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 103). Springer New York.

J.S. Cramer. (2002, November). The origins of logistic regression. Retrieved November 25, 2022, from https://papers.tinbergen.nl/02119.pdf

Kleinbaum, D. G., & Klein, M. (2012). *Survival analysis - a self-learning text* (Third). Springer. Retrieved November 25, 2022, from http://www.uop.edu.pk/ocontents/survival-analysis-self-learning-book.pdf

Kvamme, H., Borgan, O., & Scheel, I. (2019). Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*. Retrieved November 26, 2022, from https://arxiv.org/pdf/1907.00825v2.pdf

Lee, C., Zame, W., Yoon, J., & Schaar, M. v. d. (2018). DeepHit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.11842

Mata, E., Kalagasidis, A. S., & Johnsson, F. (2012, December). *Retrofitting measures for energy savings in the swedish residential building stock— assessing methodology*. https://publications.lib.chalmers.se/records/fulltext/local_123451.pdf

Mata, E., Kalagasidis, A. S., & Johnsson, F. (2018, May). *Contributions of building retrofitting in five member states to EU targets for energy savings - ScienceDirect*. Retrieved December 9, 2022, from https://www.sciencedirect.com/science/article/pii/S1364032118303575?via%3Dihub

McCartney, A., Hensman, S., & Longo, L. (2017). How short is a piece of string?: The impact of text length and text augmentation on short-text classification AccuracyText augmentation on short-text classification accuracy. Retrieved September 15, 2022, from https://arrow.tudublin.ie/cgi/viewcontent.cgi?article=1247&context=scschcomcon

McGregor, D., Palarea-Albaladejo, J., Dall, P., Hron, K., & Chastin, S. (2020). Cox regression survival analysis with compositional covariates: Application to modelling mortality risk from 24-h physical activity patterns. *Statistical Methods in Medical Research*, *29*(5), 1447–1465. https://doi.org/10.1177/0962280219864125

Minergie Schweiz. (2022a). *Baustandard minergie für modernisierungen: Der klassiker* [Minergie]. Retrieved December 1, 2022, from https://www.minergie.ch/de/standards/modernisierung/minergie/

Minergie Schweiz. (2022b). *Mit minergie zertifizieren* [Minergie]. Retrieved December 12, 2022, from https://www.minergie.ch/de/zertifizieren/minergie/

Narula, K., Chambers, J., Streicher, K. N., & Patel, M. K. (2018). Strategies for decarbonising the swiss heating system. *Energy*, *169*, 1119–1131. https://doi.org/10.1016/j.energy.2018.12.082

Nowogońska, B. (2019). The method of predicting the extent of changes in the performance characteristics of residential buildings. *Archives of Civil Engineering, Vol. 65*. https://doi.org/10.2478/ace-2019-0020

Pandey, A. (2020, April 24). *Survival analysis: Intuition & implementation in python* [Medium]. Retrieved November 25, 2022, from https://towardsdatascience.com/survival-analysis-intuition-implementation-in-python-504fde4fcf8e

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011a). *Scikit-learn: 1.1. linear models* [Journal of machine learning research]. https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011b). *Scikit-learn: Evaluating survival models with scikit-survival* [Journal of machine learning research]. https://scikit-survival.readthedocs.io/en/stable/user_guide/evaluating-survival-models.html

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011c). *Scikit-learn: Introduction to survival analysis with scikit-survival* [Journal of machine learning research]. https://scikit-survival.readthedocs.io/en/stable/user_guide/00-introduction.html

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011d). *Scikit-learn: Introduction to survival support vector machine* [Journal of machine learning research]. https://scikit-survival.readthedocs.io/en/stable/user_guide/survival-svm.html

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011e). *Scikit-learn: Working with text data* [Journal of machine learning research]. https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html

Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. Retrieved September 15, 2022,

from https://www.ra.ethz.ch/CDStore/www2008/www2008.org/papers/pdf/p91-phanA.pdf

Polly, B., Gestwick, M., Bianchi, M., Anderson, R., Horowitz, S., Christenson, C., & Judkoff, R. (2011, April 1). *Method for determining optimal residential energy efficiency retrofit packages* (NREL/TP-5500-50572). National Renewable Energy Lab. (NREL), Golden, CO (United States). https://doi.org/10.2172/1015501

Pölsterl, S. (n.d.). *Evaluating survival models* [Scikit-survival documentation]. Retrieved November 26, 2022, from https://scikit-survival.readthedocs.io/en/stable/user_guide/evaluating-survival-models.html

Prentice, R. L., Williams, B. J., & Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, *68*. Retrieved December 12, 2022, from https://doi.org/10.1093/biomet/68.2.373

Rodríguez, G. (2007). *Lecture notes on generalized linear models*. Retrieved December 2, 2022, from https://grodri.github.io/glms/notes/

Sandberg, N. h., Sartori, I., Heidrich, O., Dawson, R., Dascalaki, E., Dimitriou, S., Vimmr, T., Filippidou, F., Stegnar, G., Sijanec Zavrl, M., & Brattebo, H. (2016). Dynamic building stock modelling: Application to 11 european countries to support the energy efficiency and retrofit ambitions of the EU. *Elsevier*, *132*. https://doi.org/10.1016/j.enbuild.2016.05.100

Shi, X., Qu, T., Van Pottelbergh, G., van den Akker, M., & De Moor, B. (2022). A resampling method to improve the prognostic model of end-stage kidney disease: A better strategy for imbalanced data. *Frontiers in Medicine*, *9*. Retrieved November 24, 2022, from https://www.frontiersin.org/articles/10.3389/fmed.2022.730748

Streicher, K. N., Berger, M., Chambers, J., Schneider, S., & Patel, M. K. (2019). Combined geospatial and techno-economic analysis of deep building envelope retrofit. *Journal of Physics: Conference Series*, *1343*(1), 012028. https://doi.org/10.1088/1742-6596/1343/1/012028

Streicher, K. N., Berger, M., Panos, E., Narula, K., Soini, M., & Patel, M. (2021). Optimal building retrofit pathways considering stock dynamics and climate change impacts. *Energy policy*, *152*(112220). https://doi.org/10.1016/j.enpol.2021.112220

Streicher, K. N., Parra, D., Buerer, M. C., & Patel, M. K. (2017). Techno-economic potential of large-scale energy retrofit in the swiss residential building stock. *122*. Retrieved December 10, 2022, from https://www.sciencedirect.com/science/article/pii/S1876610217329120

Sundus, K. I., Hammo, B. H., Al-Zoubi, M. B., & Al-Omari, A. (2022). Solving the multicollinearity problem to improve the stability of machine learning algorithms applied to

a fully annotated breast cancer dataset. *Informatics in Medicine Unlocked*, *33*, 101088. https://doi.org/10.1016/j.imu.2022.101088

Swiss Federal Office for Energy SFOE. (2019). *DUREE project: Analysis of lifetimes of building elements in the literature and in renovation practices and sensitivity analyses on building LCA & LCC*. Yverdon-les-Bains. https://www.aramis.admin.ch/Default?DocumentID=67264&Load=true

Swiss Federal Office of Energy SFOE. (n.d.). *Beim Sanieren von Fördergeldern profitieren | Das Gebäudeprogramm*. Retrieved April 8, 2022, from https://www.dasgebaeudeprogramm.ch/de/das-gebaudeprogramm/forderung/

Swiss Federal Office of Energy SFOE. (2022). Energiegerecht sanieren. *Ratgeber für Bauherrschaften*. https://pubdb.bfe.admin.ch/de/publication/download/5324

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., & Wei, L. J. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, *30*(10), 1105–1117. https://doi.org/10.1002/sim.4154

Valverde-Albacete, F. J., & Peláez-Moreno, C. (2014). 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. *PLOS ONE*, *9*(1), e84217. https://doi.org/10.1371/journal.pone.0084217

Volland, B., Farsi, M., Lasvaux, S., & Padey, P. (2020, November). Too little too late: An empirical study of renovation of building elements. Retrieved November 25, 2022, from https://www5.unine.ch/RePEc/ftp/irn/pdfs/WP20-02.pdf

Weber, T. (2019, June 5). *Sanierungsdruck auf Gebäudeebene* [Swiss Energy Planning SEP]. Retrieved May 7, 2022, from https://www.swissenergyplanning.ch/post/sanierungsdruck-auf-geb%C3%A4udeebene

Weber, T. (2022, June 21). *Machine Learning-based Heat Demand Model* [SEP]. Retrieved October 16, 2022, from https://www.swissenergyplanning.ch/post/machine-learning-based-heat-demand-model

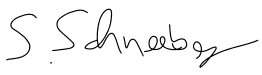
Wirth, R., & Hipp, J. (2000). Crisp-dm: Towards a standard process modell for data mining. *undefined*. Retrieved May 7, 2022, from https://www.semanticscholar.org/paper/Crisp-dm%3A-towards-a-standard-process-modell-for-Wirth-Hipp/48b9293cfd4297f855867ca278f7069abc6a9c

Wu, S. (2021, June 5). *Multi-collinearity in regression* [Medium]. Retrieved December 1, 2022, from https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea

Declaration of originality

The undersigned hereby declares that she

- wrote the work in question independently and without the help of any third party,
- has provided all the sources and cited the literature used,
- will protect the confidentiality interests of the client and respect the copyright regulations of Lucerne University of Applied Sciences and Arts.



Sarah Schneeberger

Berne, 22 December 2022

Lucerne University of Applied Sciences and Arts
Master of Science (MScIDS) in Applied Information and Data Science Lucerne

Appendix A. Data Dictionary

Table A.1. Data Dictionary: Heat Energy Consumption

Attribute	***Data Type***	***Description***
egid	Integer	Federal building identifier
year	Integer	Year of the measurement of heat energy consumption
q_use_true	Float	Heat Energy Consumption (HEC, kWh/year)

Table A.2. Data Dictionary: Building Application from Docu Media Schweiz GmbH

Attribute	***Data Type***	***Description***
baid	Integer	Unique building application ID
egid	Float	Federal building identifier
edid	Float	Federal entrance identifier
projectstreet	Object	Concerned street name of the building application
projectpostcode	Object	Concerned post code of the building application
projecttown	Object	Concerned town of the building application
projectcanton	Object	Concerned canton of the building application
projectdescription	Object	Short description of the building application
projectmaincategory	Object	Main category of the object of the building application (e.g. residential or industrial and commercial)
projectname	Object	Name of the building application
projectdate	Object	Federal entrance identifier
projectstate	Object	Date of submission of the building application
baukosten	Float	Estimated cost of the construction project applied for [1 million CHF]
projectpurpose	Object	Purpose of the construction project applied for
projectlanguage	Object	Language in which the building application is written
category_code	Object	Category codes of the object affected by the building application (one object can have several categories)
category_text	Object	Name of the category_code
category_type	Object	Each category code belongs to a category type (there are five different types in total: New construction, Reconstruction inside, Reconstruction outside, Annexe and Demolition)
ga_code	Object	Detailed information of the construction project applied for (one object can have several such ga_codes)
ga_text	Object	Name of the ga_codes
text_details	Object	Free text information about the construction project applied for

TABLE A.3. Data Dictionary: Building Characteristics

Attribute	*Data Type*	*Description*
egid	Integer	Federal building identifier
gdenr	Integer	Number of the political municipality according to the official municipal directory of Switzerland
gdename	Object	Name of the political municipality
canton	Object	Name of the canton
lat	Float	Latitude
long	Float	Longitude
geb_kategorie_gwr	Intger	Classification of buildings according to their purpose.
geb_klasse	Float	Classification of buildings, EUROSTAT, 15.10.1997
geb_klasse_name_de	Object	Name of the geb_klasse
anz_geschosse	Float	Number of floors and basements of a building, including first floor
baujahr_lower	Float	Lower limit of the construction period of the building
baujahr_upper	Float	Upper limit of the construction period of the building (if the exact year of construction is known, then baujahr_lower = baujahr_upper)
renovationsjahr_lower	Float	Lower limit of the renovation period of the building
renovationsjahr_upper	Float	Upper limit of the renovation period of the building
renovrate_neigh_current	Float	Renovation rate of the neighborhood in 2022 (calculated by geoimpact AG)
renovrate_neigh_event	Float	Renovation rate of the neighborhood in the year of the renovation of the building (calculated by geoimpact AG)
geb_flaeche2	Float	Buildings Footprint Area [m^2]
volume_egid	Float	Buildings Volume [m^3]
energy_reference_area_pred	Float	Predicted energy reference area of the building (calculated by geoimpact AG)
heating_degree_days_pred	Float	Predicted heating degree days (calculated by geoimpact AG)
anz_wohnungen_egid	Integer	Number of apartments in the building
anz_firmen_egid	Float	Number of companies in the building
heat_meas_incl_hotwater	Float	Indicates whether the energy consumption is only for heating or also for hot water
minergie	Float	Indicates whether the building complies with the Minergie standard

B.1. **GWR Construction Project Analysis.**

The GWR stores information about construction projects (renovation and new construction) in two data sources. GWR_MADD_PROJ contains general information about the construction project, like status and date information, and GWR_MADD_ARB specific information about the project, like affected building and building elements. The two datasets can be merged by the Federal Construction Project Identifier (EPROID).
Table B.1 shows the available attributes, including the number of missing values after merging the two datasets. In 41.71%, all information of the data source GWR_MADD_ARB is missing, and consequently, the information on which building is affected is missing (see attribute *EGID*). Furthermore, in more than 70%, information about the planned/realized works is missing: for example, if it is an energetic restoration (*PENSAN*), if the heating system is affected (*PHEIZSAN*), if the project includes solar heating system (*PTHERSOL*) or photovoltaics (*PPHOTSOL*). Therefore, the data from GWR_MADD_ARB are not useful but necessary since it connects the construction project to a building (*EGID*). An interesting attribute of the GWR_MADD_PROJ is the construction project description (*PBEZ*), which is available in all given projects. However, in 93.93%, the date of the completion of the project is missing, which is very high. This can be explained by the fact that completed construction projects are archived for only two years and then deactivated (Federal Statistical Office FSO, 2018).

B.2. **Comparison of Renovation Information from Different Sources.**

From Docu Media Schweiz GmbH 1'519'010 building applications recorded since 2004 are provided. In the data from RBD only 117'854 construction projects applied for are included. When comparing the attributes *EGID*, *PDATIN* from the GWR data with the corresponding attributes *egid*, *projectdate* from the building application from Docu Media Schweiz GmbH, we observe that building application tend to be registered first in the GWR and afterward from Docu Media Schweiz GmbH, which is very likely to be true (compare Figrue B.1). Therefore, we identified a building application to be the same if the attribute *EGID* matches and the *projectdate* is between *PDATIN* - 1 year and *PDATIN* + 60 days. In doing so, 41% of possible building applications that appear in both data sets remain.[11] When also including the project cost with a tolerance of ±100′000 CHF to the comparison, only 21% remain.

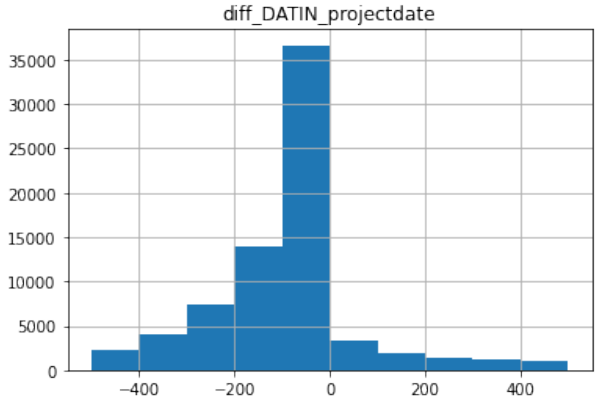


Figure B.1. Comparison of the Date of the Application

[11]Possible building applications that occur in both data sets are defined as the intersect of the two data sets by the *EGID*, and which have a *projectdate* greater or equal to 2017-01-01. This restriction is made since the RBD keeps completed construction projects only for two years in the archive (Federal Statistical Office FSO, 2018).

TABLE B.1. Missing Values of Construction Project Attributes

Source	*Attribute*	*Description*	*Number of missing values*	*%*
GWR_MADD_PROJ	EPROID	Eidgenössischer Bauprojektidentifikator	0	0
GWR_MADD_PROJ	PBDNR	Amtliche Baudossiernummer	201	0.10
GWR_MADD_PROJ	PBDNRSX	Amtliche Baudossiernummer Zusatz	4203	2.08
GWR_MADD_PROJ	PBEZ	Umschreibung Bauprojekt	0	0
GWR_MADD_PROJ	PARTBZ	Bewilligungsgrund	3989	1.98
GWR_MADD_PROJ	PTYPAG	Typ der Auftraggeber	121	0.06
GWR_MADD_PROJ	PARTBW	Art der Bauwerke	19	0.01
GWR_MADD_PROJ	PTYPBW	Typ der Bauwerke	77	0.04
GWR_MADD_PROJ	PKOST	Projektkosten total	0	0
GWR_MADD_PROJ	PDATIN	Datum Baueingabe	0	0
GWR_MADD_PROJ	PDATOK	Datum Baubewilligung	31140	15.40
GWR_MADD_PROJ	PDATBB	Datum Baubeginn	121469	60.08
GWR_MADD_PROJ	PDATBE	Datum Bauende	189896	93.93
GWR_MADD_PROJ	PDATSIST	Datum Sistierung	197542	97.71
GWR_MADD_PROJ	PDATABL	Datum Ablehnung des Baugesuchs	202156	99.99
GWR_MADD_PROJ	PDATANN	Datum Nichtrealisierung	201091	99.47
GWR_MADD_PROJ	PDATRZG	Datum Rückzug des Baugesuchs	202128	99.98
GWR_MADD_PROJ	PVBD	Voraussichtliche Baudauer	100589	49.76
GWR_MADD_PROJ	PSTAT	Status Bauprojekt	0	0
GWR_MADD_PROJ	PANZGEB	Anzahl Gebäude des Bauprojekts	84319	41,71
GWR_MADD_PROJ	PANZWHG	Anzahl Wohnungen des Bauprojekts	106786	52.82
GWR_MADD_PROJ	EGRID	Eidgenössischer Grundstücksidentifikator	192520	95.23
GWR_MADD_PROJ	BPARZ	Grundstücknummer	2059	1.02
GWR_MADD_PROJ	BGBKR	Grundbuchkreisnummer	2059	1.02
GWR_MADD_PROJ	PGDENR	Bauort (BFS-Gemeindenummer)	0	0
GWR_MADD_PROJ	GDENAME	Gemeindename	0	0
GWR_MADD_PROJ	GDEKT	Kantonskürzel	0	0
GWR_MADD_PROJ	Create_Date	Datum der Erstellung	0	0
GWR_MADD_PROJ	Update_Date	Datum der letzten Änderung	0	0
GWR_MADD_ARB	ARBID	Arbeitsidentifikator	84319	41.71
GWR_MADD_ARB	EGID	Eidgenössischer Gebäudeidentifikator	84319	41.71
GWR_MADD_ARB	PARTAB	Art der Arbeiten	84319	41.71
GWR_MADD_ARB	PENSAN	Energetische Sanierung	141816	70.15
GWR_MADD_ARB	PHEIZSAN	Sanierung des Heizsystems	142003	70.24
GWR_MADD_ARB	PINNUMB	Umbauten / Renovationen im Innenbereich	142003	70.24
GWR_MADD_ARB	PUMNUTZ	Umnutzung	142003	70.24
GWR_MADD_ARB	PERWMHZ	Beheizte Erweiterung	142003	70.24
GWR_MADD_ARB	PERWOHZ	Nicht beheizte Erweiterung	142003	70.24
GWR_MADD_ARB	PTHERSOL	Thermische Solaranlage	142003	70.24
GWR_MADD_ARB	PPHOTSOL	Photovoltaische Solaranlage	142003	70.24
GWR_MADD_ARB	PANDUMB	Andere Umbauten	141816	70.15
GWR_MADD_ARB	Create_Date (ARB)	Datum der Erstellung	84628	41.86
GWR_MADD_ARB	Update_Date (ARB)	Datum der letzten Änderung	84319	41.71

B.3. **Text analysis of Building Applications.**

Table B.2 shows the top 20 terms per language of the building application descriptions from Docu Media Schweiz GmbH. Only terms of length one to three words (so-called unigrams to 3-grams) are considered.

TABLE B.2. Top 20 (1-3)-grams per Language

	German		*French*		*Italian*	
Rank	***Word***	***Count***	***Word***	***Count***	***Word***	***Count***
1	Neubau	560004	construction	213313	nuova	30673
2	Einbau	287114	création	114809	nuovo	26320
3	Einfamilienhaus	192298	transformation	88813	abitazione	23693
4	Umbau	190844	pièces	70310	tetto	23192
5	Anbau	180384	toiture	66011	costruzione	22215
6	erhältlich	139500	places	65691	ristrutturazione	21366
7	Sanierung	120902	aménagement	60725	piano	21237
8	erweiterung	117642	chauffage	58270	m2	20707
9	Wärmepumpe	116406	habitation	56496	impianto	19935
10	Abbruch	113161	couvert	55417	casa	19736
11	sowie	109285	garage	54842	cantiere	19348
12	Angaben	106086	maison	49395	mesi	16618
13	Detail	84188	agrandissement	48569	formazione	15343
14	Garage	83841	villa	46493	unifamiliare	13046
15	Detail Angaben	83554	parc	45621	acqua	12947
16	neue	83012	pose	45082	ampliamento	12236
17	Flachdach	72986	deux	43246	copertura	12234
18	Satteldach	71078	façades	42669	termopompa	11935
19	Zimmer	70689	bâtiment	41163	legno	11528
20	erreichbar	70019	installation	40589	serramenti	11238

Appendix C. Data Preparation

C.1. Logic to Identify Unique Application Type of Building Applications.

A building application was labeled as **new construction**, if the following rule was fulfilled:

- EITHER the value of the attribute *category_type* contained the class 'Neubau' and the year of the submission of the building application was smaller or equal to the Upper limit of the construction period of the building
- OR the value of the attribute *category_type* contained the class 'Neubau' and not the class 'Umbau' and not the class 'Anbau'

A building application was labeled as **renovation**, if the following rule was fulfilled:

- the value of the attribute *category_type* contained the class 'Anbau' or 'Umbau'
- AND
  - EITHER the date of the submission of the building application minus three years was bigger than the Upper limit of the construction period of the building
  - OR the value of the attribute *category_type* does not contain the class 'Neubau' and the year of the submission of the building application was bigger than Upper limit of the construction period of the building

A building application was labeled as **demolition**, if the following rule was fulfilled:

- the value of the attribute *category_type* contained the class 'Abbruch' and not the class 'Neubau' and not the class 'Anbau' and not the class 'Umbau'

C.2. Irrelevant Categories of Building Application.

Table C.1. Identification of Irrelevant Categories

category_code	*category_text*	*irrelevant categories*
110	Lofts	
111	Einfamilienhäuser	
112	Doppel-Einfamilienhäuser	
113	Ferienhäuser	
114	Behelfswohnungen	
115	Bauernhäuser	
116	Wohnungen	
120	Mehrfamilienhäuser	
130	Terrassenhäuser	
140	Reihenhäuser	
141	Einfamilienhaus-Siedlungen	
150	Alterswohnungen, Alterssiedlungen	
160	Alterswohnheime	
170	Kinder- und Jugendheime	
180	Studenten- und Lehrlingswohnheime	
210	Kinderhorte und Kindergärten	
220	Primar- und Sekundarschulen	
230	Berufs- und höhere Fachschulen	
240	Mittelschulen und Gymnasien	
250	Heilpädagogische Schulen/Sonderschulen	
260	Hochschulen und Universitäten	
270	Bibliotheken und Staatsarchive	
280	Forschungsinstitute	
310	Lagerhallen	
320	Mehrgeschossige Lagerbauten	
330	Mechanisierte Lager und Kühllager	
340	Silobauten und Behälter	
350	Verteilzentralen	
360	Industriehallen	
370	Industrielle Produktionsbauten	
380	Betriebs- und Gewerbebauten	

category_code	*category_text*	*irrelevant categories*
381	Autowerkstätten	
382	Atelier und Studio	
410	Schuppen und Hütten	
411	Gartenhäuser / Pavillons	
420	Futterlagerräume, Treibhäuser und Silobauten	
430	Stallungen und landwirtschaftliche Produktionsanlagen	
440	Tierheime und Veterinärstationen	
450	Tierspitäler	
460	Schlachthöfe	
480	Jauchegrube	x
510	Heizzentralen, Fernwärmeanlagen und Kraftwerkbauten	
520	Wasseraufbereitungsanlagen	
530	Kehrichtverbrennungs- und Wiederaufbereitungsanlagen	
531	Deponien	x
540	Tankanlagen, Tankstellen	
550	Masten, Türme	x
560	Wärme- und Kälteverteilanlagen	x
570	Elektrische Verteilanlagen	x
580	Verteilanlagen für Trinkwasser	x
590	Autogewerbe	
610	Ladenbauten	
620	Warenhäuser und Einkaufszentren	
630	Bürobauten mit einfachen Anforderungen	
640	Bürobauten mit erhöhten Anforderungen	
650	Verwaltungsgebäude und Rechenzentren	
660	Banken, Postgebäude und Fernmeldegebäude	
670	Gemeindehäuser, Rathäuser und Regierungsgebäude	
710	Gerichtsgebäude	
720	Polizeieinsatzgebäude und Untersuchungsgefängnisse	
730	Strafvollzugsanstalten	
740	Wiedereingliederungsstätten	
810	Arztpraxen und Ärztehäuser	
820	Krankenhäuser	
830	Universitätskliniken	
840	Pflegeheime, Sanatorien und Rehabilitationszentren	
850	Heilbäder und Spezialinstitute	
860	Tagesheime und geschützte Werkstätten	
1110	Restaurationsbetriebe	
1120	Hotel- und Motelbauten	
1130	Kantinen	
1140	Herbergen, Jugendherbergen und Massenunterkünfte	
1150	Raststätten, Cafeterias, Tea-Rooms und Bars	
1160	Klubhütten	
1170	Berghäuser	
1180	Campinganlagen	
1210	Sportanlagen, Turn- und Mehrzweckanlagen	
1220	Tribünenbauten und Garderobengebäude	
1230	Hallen- und Freibäder	
1231	Swimmingpools, Jacuzzi	x
1240	Reithallen	
1250	Bootshäuser	
1260	Freizeitzentren und Jugendhäuser	
1270	Aussenanlagen, Kinderspielplätze und Parkanlagen	x
1280	Zoologische und botanische Gärten, Gewächshäuser	x
1290	Fitnesscenter/-raum	
1310	Parkhäuser und Einstellhallen	
1311	Parkplätze	x
1312	Tiefgaragen und Unterniveaugaragen	x
1313	Garagen / Fertiggaragen	
1314	Carports und Abstellplätze	x
1320	Strassenverkehrsgebäude	
1330	Werkhöfe	
1340	Busbahnhöfe, Zollanlagen und Wartehallen mit Diensträumen	
1350	Bahnhöfe und Bahnbetriebsbauten, Seilbahnstationen	
1360	Flughafenbauten	
1370	Post- und Logistikterminale	
1410	Kasernen	
1420	Zeughäuser	
1430	Öffentliche Zivilschutzanlagen	
1440	Sanitätsposten und Sanitätshilfsstellen	
1450	Geschützte Operationsstellen und Notspitäler	
1460	Zivilschutz-Ausbildungszentren	
1470	Feuerwehrgebäude	
1480	Militäranlagen und militärische Schutzanlagen	
1510	Kirchen und Kapellen	
1520	Kirchgemeindehäuser	
1530	Friedhofanlagen	
1540	Abdankungshallen	
1550	Krematorien	
1560	Klöster	
1570	Burgen & Schlösser	
1571	Denkmäler, Kunstbauten	
1610	Ausstellungsbauten	
1620	Museen und Kunstgalerien	
1630	Wohlfahrtshäuser, Klubhäuser und Kulturzentren	
1640	Konzertbauten und Theaterbauten	
1650	Musikpavillons	
1660	Kino-, Diskothek- und Saalbauten	

category_code	*category_text*	*irrelevant categories*
1670	Kongresshäuser und Festhallen	
1680	Radio-, Fernseh- und Filmstudios	
1690	Casino	
1710	Wintergarten	
1770	Öffentliche WC-Anlagen	
1780	Personenunterstände	x
2110	Strassen	x
2120	Autobahnen	x
2210	Eisenbahnen	x
2220	Magnetbahnen	x
2310	Strassenbahnen	x
2410	Waldwege	x
2420	Flurwege	x
2510	Pisten	x
2520	Rollwege	x
2610	Freispiegelkanäle, nicht befahrbar	x
2620	Schiffahrtskanäle	x
2630	Binnenhafenanlagen	x
2640	Werftanlagen	
2710	Erddämme	x
3110	Kanalisationen innerorts	x
3120	Kanalisationen ausserorts	x
3130	Kanalisationen für Industrieanlagen	x
3140	Drainagen	x
3150	Bauwerke für Entwässerungen	x
3210	Hochspannungsleitungen	x
3220	Signal- und Fernmeldeleitungen	x
3230	Gas- und Wasserverteilleitungen	x
3240	Fernwärme-Verteilleitungen	
3250	Pipelines	x
3260	Zisternen	x
3310	Grundwasserschutz innerorts	x
3320	Grundwasserschutz ausserorts	x
3410	Inertstoffdeponien	x
3420	Reststoffdeponien	x
3430	Reaktordeponien	x
3440	Sondermülldeponien	x
3510	Leitdämme	x
3520	Abfangdämme	x
3530	Deiche und Dammbauten	x
3610	Lärmschutzmassnahmen	x
4110	Fussgänger- und Radfahrerbrücken	x
4120	Strassenverkehrsbrücken	x
4130	Eisenbahnbrücken	x
4210	Talsperren	x
4310	Stützmauern	x
4410	Lawinenverbauungen	x
4510	Tagbautunnels	x
4520	Galerien	x
4610	Wildbachverbauung	x
4620	Ufer- und Sohlensicherungen	x
4630	Fischteiche	x
4640	Flussregulierungen	x
4810	Schleusenanlagen	x
4820	Wehranlagen	x
4830	Einlaufbauwerke	x
4840	Mündungsbauwerke	x
4850	Uferbefestigungen	x
5110	Mechanische Abwasserreinigungsanlagen	x
5120	Biologische und chemische Abwasserreinigungsanlagen	x
5210	Regenrückhaltebecken	x
5220	Hochwasserrückhaltebecken	x
5230	Talsperren	x
5310	Düker	x
6110	Sessellifte und Gondelbahnen	x
6120	Zahnradbahnen	x
6130	Standseilbahnen	x
6140	Luftseilbahnen	x
6210	Richtstrahlanlagen	x
6310	Antennanlagen	x
6410	Windkraftanlagen	x
6510	Hochspannungs-Trafostationen	x
6520	Unterwerke	
6610	Silos	x
6620	Wassertürme	x
7110	Strassentunnel	x
7120	Bahntunnel	x
7210	Abwasserstollen	x
7220	Druckstollen	x
7310	Kavernen	x
7410	Vertikale Schächte	x
7420	Geneigte Schächte	x
7510	Portale	x
7520	Lüftungsbauwerke	x
9051	Hochbau	
9052	Tiefbau	
9054	Lieferungen aller Art	x

D.1. **Rules to Classify Building Application by Preexisting Categories.**

Table D.1 contains the assignment of ga_code-classes to a structural measure (building envelope, building technology), to sub-groups of the structural measure (window, roof, facade, insulation, heating and hot water, ventilation and air conditioning, solar heat, solar power), and to the dummy variables possible energetic, energetic and minergie. With this assignment and the application type (see Appendix C.1), I classified the building application. And if a building application was classified as renovation, energetic, and building envelope, we identified and classified it as energetic restoration.

Table D.1. Aggregations of Preexisting Classes of the Attribute ga_code

ga_code	*ga_text*	*Structural measures*	*Struct. meas. sub-classes*	*possible energetic*	*energetic*	*minergie*
100	Dächer ohne Detailangaben	Building envelope	Roof	x		
101	Flachdach	Building envelope	Roof	x		
102	Schrägdach	Building envelope	Roof	x		
104	Dachumbau, Dachausbau	Building envelope	Roof	x		
105	Dachflächenfenster	Building envelope	Roof			
106	Lukarnen	Building envelope	Roof			
107	Vordach	Building envelope	Roof			
108	Lichtkuppel	Building envelope	Roof			
109	Minergie-Standard	Building envelope	Roof, Insulation	x	x	x
150	Dacheindeckungen ohne Detailangaben	Building envelope	Roof			
151	Ziegel	Building envelope	Roof	x		
152	Faserzement	Building envelope	Roof	x		
153	Naturstein, Schiefer	Building envelope	Roof			
154	Dachbegrünungen	Building envelope	Roof, Insulation	x	x	
155	Bitumen, Teer	Building envelope	Roof			
156	Blech / Metall / Panel	Building envelope	Roof			
200	Fassaden ohne Detailangaben	Building envelope	Facade	x		
201	Metall, Stahl, Leichtmetall	Building envelope	Facade	x		
202	Holz	Building envelope	Facade	x		
203	Naturstein	Building envelope	Facade	x		
204	Glas	Building envelope	Facade	x		
205	Mauerwerk verputzt	Building envelope	Facade	x		
206	Fassadenelemente: Beton, Leichtbeton, Kunststein	Building envelope	Facade	x		
207	vorgehängte, hinterlüftete Fassaden	Building envelope	Facade, Insulation	x	x	
208	Faserzementplatten	Building envelope	Facade	x		
209	Keramik	Building envelope	Facade	x		
210	Sichtmauerwerk	Building envelope	Facade	x		
211	Sandwich-Panel	Building envelope	Facade, Insulation	x	x	
212	Sichtbeton	Building envelope	Facade	x		
214	Kompaktfassaden	Building envelope	Facade, Insulation	x	x	
215	Minergie-Standard	Building envelope	Facade, Insulation	x	x	x
300	Fenster, Fenstertüren ohne Detailangaben	Building envelope	Window	x		
301	Metall-, Leichtmetallfenster	Building envelope	Window	x	x	
302	Holzfenster	Building envelope	Window	x	x	
303	Kunststofffenster	Building envelope	Window	x	x	
304	Isolier-, Wärmedämm-, Schalldämmglas	Building envelope	Window, Insulation	x	x	
305	Balkon-, Terrassenverglasung	Building envelope	Window			
306	Holz/Metallfenster	Building envelope	Window	x	x	
307	Schaufensteranlagen	Building envelope	Window			
313	Minergie-Standard	Building envelope	Window Insulation	x	x	x
400	Tragkonstruktion ohne Detailangaben					
401	Holz					
402	Beton					
403	Backstein					
404	Porenbetonstein					
405	Kalksandstein					
406	Skelettbau (Beton, Stahl, Holz)					
407	Stahl					
408	Zweischalenmauerwerk					
409	Sichtmauerwerk					
410	Einsteinmauerwerk					
500	Heizung ohne	Building technology	Heating and			

ga_code	*ga_text*	*Structural measures*	*Struct. meas. sub-classes*	*possible energetic*	*energetic*	*minergie*
	Detailangaben		hot water			
501	Oelheizung	Building technology	Heating and hot water			
502	Gasheizung	Building technology	Heating and hot water			
503	Fernwärme	Building technology	Heating and hot water	x	x	
504	Wärmepumpen	Building technology	Heating and hot water	x	x	
505	Elektroheizung	Building technology	Heating and hot water			
506	Solarheizsysteme	Building technology	Solar heat	x	x	
507	Holzheizung	Building technology	Heating and hot water	x	x	
508	Cheminées, Cheminéeöfen	Building technology	Heating and hot water	x	x	
509	Bodenheizung	Building technology	Heating and hot water	x		
510	Heizkörper: Radiatoren, Heizwände	Building technology	Heating and hot water	x		
511	Geothermie, Erdwärmesonden/-kollektoren	Building technology	Heating and hot water	x	x	
512	Holzschnitzelheizung	Building technology	Heating and hot water	x	x	
513	Pelletheizung	Building technology	Heating and hot water	x	x	
514	Kontrollierte Raumbelüftung, Komfortlüftung	Building technology	Ventilation and air conditioning	x x	x x	
600	Boden ohne Detailangaben					
601	Unterlagsboden					
602	Kunststeinboden					
603	Parkettboden / Korkboden					
604	Linoleumboden, Kunststoffboden					
605	Textilboden					
606	Keramikboden					
607	Holzboden					
608	Betonboden					
609	Doppelboden, Hohlraumboden					
610	Natursteinboden					
611	Laminatboden					
612	Industrieboden, fugenlos					
700	Dämm- und Abdichtung ohne Detailangaben	Building envelope	Insulation	x	x	
701	InnenwärmeInsulation	Building envelope	Insulation	x	x	
702	AussenwärmeInsulation	Building envelope	Insulation	x	x	
706	ZwischenwärmeInsulation	Building envelope	Insulation	x	x	
707	PerimeterwärmeInsulation	Building envelope	Insulation	x	x	
714	Minergie-Standard	Building envelope	Insulation	x	x	x
800	Innenausbau ohne Detailangaben					
801	Haushaltsküchen					
802	Gewerbliche Küchen, Grossküchen					
803	Bad, WC, Dusche					
804	Laboreinrichtungen					
805	Keller / Hobbyraum					
806	Sauna / Dampfbad					
807	Wohnung					
808	abgehängte Decken					
809	Treppen					
810	Gegensprechanlage					
811	Maler-/Gipserarbeiten					
812	Waschküche					
813	Kaminsysteme					
814	Sanitäranlagen					
815	versetzbare Trennwände					
816	Einbauschränke					
817	Whirlpool / Jacuzzi					
818	Trockenbausysteme					
901	Klima	Building technology	Ventilation and air conditioning	x x	x	
902	Förderanlage					
903	Sonnen- / Wetterschutz	Building envelope	Window	x		
904	Gebäudeautomation					
905	Sicherheitstechnik					
906	Garagentore					
907	Umgebungsgestaltung					
908	Kälteanlagen	Building technology	Ventilation and air conditioning	x x	x	
909	Tankanlagen (Heizbereich)					
910	Terrassen / Balkone					
917	Lüftung	Building technology	Ventilation and air conditioning	x x	x	
918	Biotop / Teiche					

ga_code	*ga_text*	*Structural measures*	*Struct. meas. sub-classes*	*possible energetic*	*energetic*	*minergie*
919	Wohntüren					
920	Spezialtüren (H/G/I)					
921	Pergola					
923	Industrietore					
925	Briefkästen					
926	Aussenbeleuchtung					
927	Bewässerungsanlagen					
928	kontrolliertes Parksystem					
1000	Elektrizität ohne Detailangaben					
1001	230V					
1002	400V					
1003	500V					
1004	Solarenergie	Building technology	Solar power	x	x	

D.2. **Rules to Classify Building Application by Text Data.**

The structural measures (or word part of them) given in Table D.2 were used to search for similar words in the building applications. The resulting word list was manually checked, and wrongly selected words were deleted (such as sound insulation). Table D.3 contains the resulting dictionary, which was finally used to assign the corresponding structural measures to the building applications. More precisely, if the text data of a building application contained a word of the word list, the corresponding Word Part was added as a key to the building application. And according to Table D.2, the corresponding Sub-group and Group were assigned to the building application. Building applications of the application type renovation were classified as energetic restoration if they were assigned to the group building envelope or general.

TABLE D.2. Structural Overview of Word Parts Related to Energetic Renovation Activities

Structural measures		
Group	***Sub-group***	***word part***
Building envelope	Insulation	dämmung/dämung/dammung/dmmung
		fassade
		vip
		isolation/isolierung
		dachbegrü
	Window	fensterer
		dreifach
		verglasung
		sonnenschutz
Building technology	Heat and hot water	wärmepumpe
		erdwärme
		geotherm/thermie
		fernwärme
		holzfeuerung/holzheizung
		holzschnitzelheizung
		pelletheizung
	Solar (heat and power)	solar
		sonnenkol
		photovoltaik
	Ventilation and air conditioning	komfortlüftung
		kaskadenlüftung
		verbundlüftung
		geocooling
		rreecooling
General		energetisch
	minergie	minergie

TABLE D.3. Dictionary of Words Related to Energetic Renovation Activities

word part	*word list*
dämmung	[aussendämmung, wärmedämmung, aussenwärmedämmung, fassadendämmung, dämmung, aussendämmungen, aussenwanddämmung, fassadenaussenwärmedämmung, fassadenwärmedämmung, kompaktdämmung, zwischenwärmedämmung, dachflächendämmung, dachdämmung, hausdachdämmung, eärmedämmung, wäremdämmung, aussenwäremdämmung, nachdämmung, aussenwandwärmedämmung, aussenwarmedämmung, aufdachdämmung, dachaussenwärmedämmung, wärmedämmungssanierung, aussenwärmedämmungsverputz, warmedämmung, wärmedämmungsdach, wärmedämmungen, innendämmung, nachdämmungen, cellulosedämmung, ausssendämmung, aussemdämmung, volldämmung, zwischendämmung, kellerdämmung, fassadenaussendämmung, flachdachdämmung, kerndämmung, gebäudedämmung, aussenwärrmedämmung, aussendämmungh, wämedämmung, aussenwärmedämmunge, zellulosedämmung, aussenwärmedämmungen, aussenwärmdämmung, dachwärmedämmung, estrichdämmung, dämmungsmassnahmen, innendämmungen, dämmungen, vakuumdämmung, perimeterdämmung, holzfaserdämmung, dachdeckendämmung, wärmedämmungsverbundsystem, wändewärmedämmung, innenwärmedämmung, aussenwärmeedämmung, kompaktaussendämmung, deckendämmung, teilwärmedämmung, aussenperimeterdämmung, schafwolldämmung, fasadendämmung, sockeldämmung, zusatzdämmung, bodendämmung, aussenwärmedämmungund, ausswärmedämmung, kompaktwärmedämmung, aussedämmung, aufsparrdämmung, fassdendämmung, estrichbodendämmung, wanddämmung, verputzteraussendämmung, untersichtdämmung, zwischenwärmdämmung, mineralfaserdämmung, perimeterwärmedämmung, sparrendämmung, mineralwilldämmung, zwischensparrendämmung, zwiswchenwärmedämmung, zwischenwärnedämmung, gebäudewärmedämmung, kellerdeckendämmung, zwischenparrendämmung, einbaublasdämmung, untersparrendämmung, ausendämmung, wqärmedämmung, aussenwäredämmung, hanfkalkdämmung, innenwärmedämmungen, zwischenwämedämmung, aussenwärnedämmung, betonkerndämmung, ausenwärnmedämmung, zwischdämmung, ausenwärmedämmung, faserdämmung, zwischenwärmnedämmung, aussenwärmnedämmung, zwiscdhenwärmedämmung, zwishenwärmedämmung, wärmdämmung, mitaussendämmung, bodenplattendämmung, zwischenwärdämmung, aussenwärmecdämmung, dämmungssanierung, auissenwärmedämmung, aufsparrendämmung, zwqischenwärmedämmung, aussemwärmedämmung, zwischwenwärmedämmung, inndenwärmedämmung, neudämmung, wändedämmung, aussenndämmung, zwischenwärnmedämmung, aussnwärmedämmung, aussenwärmmedämmung, ausssenwärmedämmung, holzdämmung, aussenwäemedämmung, aufdämmung, fassadendämmungen, zwiscchenwärmedämmung, aerogeldämmung, ausserwärmedämmung, zwischenwärmedämmungh, zwischzenwärmedämmung, zwisczhenwärmedämmung, aufsparrwärmedämmung, zwischenwärmedämmungl, fassädendämmung, ausswenwärmedämmung, zweischenwärmedämmung, mineraldämmung, kernwärmedämmung, ausswendämmung, aussenwämedämmung, hochleistungsdämmung, wassadendämmung, hohlraumdämmung, mineralwolldämmung, multipordämmung, kompatkfassadendämmung, zwischwnwärmedämmung, zwisvchenwärmedämmung, einblasdämmung, aussenwäürmedämmung, aussenwärmedämmungh, dachdämmungen, aussenwqärmedämmung, aussenqwärmedämmung, sussenwärmedämmung, fasssadendämmung, zwiischenwärmedämmung, zwischenwärmedämmungi, aussendwärmedämmung, dachaufdämmung, erdwärmedämmung, aqussenwärmedämmung, zweischalenwärdämmung, gebäudehüllendämmung, zwiscvhenwärmedämmung, aussewärmedämmung, aussendärmedämmung, aussebwärmedämmung, glasfaserdämmung, fassadedämmung, dachinnendämmung, kompakwärmedämmung, ausse4nwärmedämmung, assenwärmedämmung, sandwichdämmung, aussenwändeaufdämmung, zwischnenwärmedämmung, innenwärmdämmung, jaussenwärmedämmung, aussenwüärmedämmung, estrichbodedendämmung, zwischebwärmedämmung, holzfaserplattendämmung, aussenweärmedämmung]
fassade	[fassadenisolation, fassadenisolierung, fassadendammung, fassadenisolationen, fassadenaussenwärmedämmung, fassadenwärmedämmung, fassadenisoliation, fassadenaussenisolation, fassadenisol, kompaktfassade, sandwichpanéelfassade, sandwichfassade, kompaktfassaden, sandwichpaneelfassade, sandwichpanelfassade, solarglasfassadenelemente, dämmbetonfassade, fassadenaussendämmung, fassadenisolaltion, sandwichbetonfassade, sandwichfassaden, fassadeisolation, fassadenhausisolationsverkleidung, sandwichblechfassade, fassadenislation, aussenfassadendämung, thermfassade, dämmklinkerfassade, sandwichelementenfassadenbau, aussenisolationsfassade, eternitkompaktfassade, sandwichpaneelenfassade, fassadendämmplatten, fassadenbepflanzung, kopmaktfassade, fassadenisolationsverkleidung, dämmputzfassade, komaktfassade, fassadenbewuchs, photovoltaikfassade, dämmfassade, sandwichpaneefassade, kompacktfassade, kompaktfassadenanteil, fassadenisulation, fassadenämmung, kampaktfassaden, kompakfassade, kompakfassaden, begrünterfassade, wärmekompaktfassade, kompaktgfassade, fassadendämmungen, fassadeninsolation, photovoltaikfassaden, sandwichfassadenelemente, kompkatfassade, kompatkfassadendämmung, sandwichpanéefassade, metallsandwichpanéelfassade, kompaktsteinfassade, kimpaktfassaden, fassadenbegrünung, fassadendmmung, sandwichpaneelfassaden, fassadedämmung, komopaktfassaden, kompaktgfassaden, kombaktfassade, fassadenphotovoltaikanlage, kopaktfassaden, konmpaktfassaden, fassadenbegrünungen, solarfassade, kompaktfassadem, kompktfassade]

word part	*word list*
solar	[solaranlage, solarzellen, solarheizung, solarpanel, solarzelle, solarpanele, solarkollektoren, solarlift, solarlicht, solarzellenanlage, solarsiedlung, dachsolarzellen, solaranlagen, solarmatten, solarenergie, solar, solarbeheizt, solargenerator, solarnutzung, solarkollektoranlage, solarhausplexus, solarthermieanlage, solarspeicher, solarnalage, solarsystem, solarpaneelen, solarlichtanlage, solarstromanlage, solarkollaktoren, solarzusatzheizung, solarkollektor, solargeneratoren, solarstromanalge, solarmodule, solarunterstützung, solarpanels, solarelemente, solarthermische, solarenergieanlage, solareinspeisung, solarpaneels, solardächer, solardach, dachsolarsystem, solarstromanlagen, solarabsorber, solarkollektorenanlage, solarmodulen, solarfläche, solarthermie, solaranalge, solarenergieplatte, solardachsteine, solarstrom, solarhäuser, dachsolaranlage, solarlaminate, solarabsorberplatten, solarenergieerzeugnissanlage, solarpaneele, solarschläuche, solaranlge, solaranalage, solarzellenpannelen, solarzellenaufbau, solarglasfassadenelemente, solarhaus, solaranage, kompaktsolaranlage, thermosolaranlage, solarpaneel, solarwarmwasseraufbereitungsanlage, solarboiler, solarpannelen, solarkollektorfeldes, solarröhrenkollektoren, solarnlage, solarenergieplatten, solarenergieanlagen, thermosolarzellen, solarwasseraufbereitungsanlage, solareindeckung, solarbedachung, solarpanelen, solarwasseraufbereitung, solarflächen, solarwarmwasseraufbereitung, solarkollektors, solarspeicherheizung, solarenergianlage, solarananlage, solarmodul, solarwand, solarmatte, aussensolaranlage, solarstation, solarunterstützten, solarkollektoen, solarfeld, solarenergieanlag, solartermie, solarpaneeleindeckung, solarstromzellen, solarziegeleindeckung, solarpalele, solardachfläche, solarkraftwerk, solarpanellen, solarenergeianlage, solarstromanlge, solarnalge, solarveranda, solardusche, solarthermischen, solaranlag, solarnetzeinspeisung, solarkollektoranlagen, modulsolarkollektoren, solarenergieunterstützung, solarenergiepaneelen, solarkraftwerke, solarzellendach, solarglas, solarcarport, solarbeheizten, solaranlange, solarlkollektoren, solararanlage, solareinbau, solarnetz, solarwärmeanlage, solarablage, solarvoltaikanlage, solaranlafe, solardachziegeln, solarwärme, solaranklage, solarkraftwerkes, solarkoellektoren, solarkollektorenaufbau, solarkollektorenbedachung, thermosolar, solartermische, solarstromanlageauf, solarkollektorenmontage, solarabdeckung, solarkraftwerks, solaranlagenaufbau, solarstrommodule, solarkollektorfeld, solarkollektore, solarspeichersysteme, solarplatten, solarglasziegel, solarelektrische, solarziegel, solarwasser, solarkraftwerkanlagen, solarwassererwärmung, solarpaneelanlage, solarpanell, solarziegeln, solarhermie, solartanlage, solarange, solarthermikkollektoren, solarflachkollektoren, warmwassersolaranlage, solartechnik, solarspeicherraum, solarthermieanlagen, solarthermiepaneelen, solartherme, solarwärmeverbund, solarverbund, solardachsystem, solarpanellanlage, solarthermikanlage, solarheizsyteme, solarthermiermodule, solaranlagee, solarkrafwerk, warmwassersolar, solarvoltaik, solarnanlage, solarheizsystem, solarwälrmeanlage, solarheizsysteme, solarnanlagen, solarheizsitem, solarblume, solarenegie, gartensolaranlage, solarpanelle, indachsolaranlage, solarmodulanlagen, doppelsolarcarport, solarüberdachung, luftsolarpanels, solarstromspeicher, solarpanéelen, solarpaneleindeckung, solarelementen, solarwärmekollektoren, solartheermie, solartracker, solargeländer, solarhybridanlage, solarapanels, solaranl, solarbetrieben, solarcarports, solarthermik, solarbalkongeländer, kleinsolaranlage, photovoltaiksolarmodule, solarfaltdach, solarpenels, solarbank, solarleuchten, solarpanéele, solarpanéeleindeckung, photovoltaiksolarsystemanlage, solaraktive, solarmodulverkleidung, solargewinn, thermischsolaranlage, fassadensolaranlage, solarüberdachungen, solarschindeln, solarfaltdachanlage, solartherm, solarzaun, solarkraftwerkteil, solarpflastern, solarzäunen, solarodule, solarthermischer, solarpark, solarterrasse, solarthermianlage, solarfassade, solaranöage, solaranlagenm, aufdachsolaranlage, solarpanéels, solaragnlage]
fernwärme	[fernwärmeleitungen, fernwärme, fernwärmeleitung, fernwärmenetz, fernwärmeverbund, fernwärmeanlage, fernwärmegenossenschaft, fernwärmewerk, fernwärmerohre, fernwärmeheizung, fernwärmeversorgung, fernwärmeverbundes, fernwärmeeitung, fernwärmeprojekt, fernwärmeanschluss, fernwärmeleitungsnetz, fernwärmeleitungsnetzes, fernwärmenetzes, fernwärmeversorgungsnetz, fernwärmeerzeugung, holzfernwärme, fernwärmeversorungsnetz, fernwärmeheileitung, fernwärmeast, fernwärmeversorgungsanlage, fernwärmezentralen, fernwärmehauptleitung, fernwärmeerschliessung, fernwärmesystem, fernwärmeheiznetz, fernwärmetrasse, fernwärmetrassee, fernwärmerohr, fernwärmeverteilung, fernwärmemeldeleitung, fernwärmeheizzentrale, fernwärmeversorgungsleitung, fernwärmenetzerweiterung, fernwärmeleitungsbau, fernwärmenenetz, fernwärmenetzt, fernwärmerohren, fernwärmeerschliessungsnetz, fernwärmeund, fernwärmeleitungsnetzleitungen, fernwärmed, fernwärmeleitungsanschluss, fernwärmenetzen, fernwärmeverteilleitung, fernwärmeverteileitung, fernwärmestation, fernwärmeverteilleitungen, fernwärmeanscluss, fernwärmemetz, fernwärmel, fernwärmeversorgungsleitungen, fernwärmespeicher, fernwärmeleitungnetz, fernwärmenetzleitung, fernwärmezuleitung, fernwärmen, fernwärmeleitungsrohr, fernwärmenutz, fernwärmeverbung, fernwärmeverbundleitung, fernwärmepumpe, fernwärmetransportleitung, fernwärmeanschuss, fernwärmeleitungsnetztes, fernwärmenetzten, fernwärmeerschliiessung, fernwärmeleitungenin, fernwärmeleitungleitungen, fernwärmeinstallation, fernwärmeleirung]

word part	*word list*
wärmepumpe	[wärmepumpe, luftwärmepumpe, wärmepumpenheizungen, erdsondenwärmepumpenanlage, wasserwärmepumpe, wärmepumpenheizung, wärmepumpenanlage, wärmepumpen, grundwasserwärmepumpe, erdwärmepumpenanlage, luftwärmepumpenheizung, erdsondenwärmepumpe, wärmepumpeanlage, wärmepumpengerät, aussenwärmepumpe, wärmepumpenleitung, erdwärmepumpe, gaswärmepumpe, wärmepumpenraum, grundwasserwärmepumpenanlage, wärmepumpensystem, luftwärmepumpenanlage, erdsondenwärmepumpen, wärmepumpenheinzung, wärmepumpenanlagen, aussenwärmepumpenheizung, wärmepumpeheizung, kleinwärmepumpe, wasserwärmepumpen, erdsondenwärmepumpenanlagen, aussenwärmepumpen, erdsonenwärmepumpenanlage, luftwärmepumpengerät, wärmepumpenanalge, wärmepumpenaussengerät, wärmepumpentyp, erdswondenwärmepumpe, wärmepumpenhaus, wärmepumpensystems, stromwärmepumpe, aussenwärmepumpenanlage, wasserwärmepumpenanlage, luftwasserwärmepumpe, wärmepumpenenlage, erdsondenwärmepumpenalge, splittwärmepumpe, wärmepumpemanlage, sonnenkollektorenwärmepumpe, luftwasswerwärmepumpe, umluftwärmepumpe, elektrowärmepumpe, luftwärmepumpen, wärmepumpem, ersondenwärmepumpe, wärmepumpeheizungseinbau, wärmepumpenalange, wärmepumpehiezung, wärmepumpenagregat, erdsondenwärmepumpenanalagen, einbauwärmepumpe, wärmepumpenlanlage, wärmepumpenbauwerk, grundeasserwärmepumpe, wärmepumpenheizungsanlage, wärmepumpenheizugsanlage, splitwärmepumpe, wärmepumpenzentrale, grundwasserwärmepumpen, wärmepumpenheizanlage, luftwasswerwärmepumpen, luftwasserwärmepumpenanlage, luftwassserwärmepumpe, luftwasssrwärmepumpe, wärmepumpenspeicher, wärmepumpemit, lüftwärmepumpe, wärmepumpenheizungmit, wärmepumpenkühler, wärmepumpensanierung, aussenluftwärmepumpe, aussenluftwasserwärmepumpe, solewasserwärmepumpe, wassserwärmepumpe, luftwasserwärmepumpen, wwärmepumpe, erdsondenwärmepumpenheizung, wärmepumpenaufstellung, erdsondewärmepumpe, wärmepumpeund, wasserwasserwärmepumpe, aussernwärmepumpe, erdsondewärmepumpenanlagen, innenwärmepumpe, luftwassersplitwärmepumpe, luftwasserwärmepumpenheizung, wärmepumpenboiler, wärmepumpeanla, heizungluftwasserwärmepumpe, luftwarmwärmepumpe, wärmepumpeluft, luftwasserwärmepumpeheizung, splitluftwärmepumpe, installationluftwasserwärmepumpe, solewasserwärmepumpenanlage, luftwasserwärmepumpeanlage, vonwärmepumpen, wärmepumpengeräte, wärmepumpenklimageräte, luftwärmepumpenanlagen, wasserwärmepumpenanlagen, wärmepumpenbolier, luftwaserwärmepumpe, luftluftwärmepumpe, solewärmepumpe, luftwassersplittwärmepumpe, wassertwärmepumpe, solwasserwärmepumpe, erdsondenwärmepumpemanlage, erdwärmesondenwärmepumpe, wärmepumper, erdsonbenwärmepumpenanlage, seewasserwärmepumpe, luftewasserwärmepumpe, splitluftwasserwärmepumpe, erdwärmepumpenheizung, luftwasserrwärmepumpe, solewasserwärmepumpen, aufstellenwärmepumpe, erdsodenwärmepumpenanlage, luftwassewärmepumpe, kuftwasserwärmepumpen, lufwasserwärmepumpe, wärmepumpeboiler, zwischenwärmepumpe, wärmepumpenund, slewasserwärmepumpe, fernwärmepumpe, erstellenaussenwärmepumpe, wärmepumpenfunktion, luftwasserwärmepumpem, erdsondenbohrungwärmepumpe, luftwasserewärmepumpe, luiftwasserwärmepumpe, luftwassrwärmepumpe, lufzwasserwärmepumpe, liftwasserwärmepumpe, luftwaseerwärmepumpe, wärmepumpee, wärmepumpenehizung, liuftwasserwärmepumpe, wärmepumpeaussen, wärmepumpenalage, luftwasseerwärmepumpe, installationwärmepumpe, erwärmepumpenanlage, wärmepumpeen, einerwärmepumpe, solewassserwärmepumpe, luftwsserwärmepumpe, installationluftasserwärmepumpe, luftwassertwärmepumpe, lufztwasserwärmepumpe, erdsondesolewasserwärmepumpe, léuftwasserwärmepumpe, luftwasserwärmepumpezweifamilienhaus]
erdwärme	[erdwärmesonden, erdwärme, erdwärmesonde, erdwärmepumpenanlage, erdwärmenutzung, erdwärmepumpe, erdwärmeentzug, erdwärmesondenanlage, erdwärmesondenbohrung, erdwärmeheizung, werdwärmeentzug, erdwärmekörben, erdwärmesone, erdwärmesondenbohnung, erdwärmesonen, erdwärmesondenbohrungen, erdwärmeanlage, erdwärmebohrung, erdwärmesenden, erdwärmekörbe, erdwärmesondenhiezung, erdwärmensonden, erdwärmesondenheizung, erdwärmekollektoren, miterdwärmesonde, erdwärmebohrungen, erdwärmesondernbohrung, erdwärmesondenpumpenanlage, erdwärmesondenbühungen, erdwärmesondenbohgung, erdwärmeswonde, erdwärmesondenwärmepumpe, erdwärmesonebohrung, erdwärmesondenbohung, erdwärmepumpenheizung, erdwärmesonenbohung, erdwärmedämmung, erdwärmesänden, erdwärmesondengohrung, erdwärmesnden, erdwärmesondenbohnrung, erdwärmesondenbohrungne, erdwärmesoden]
isolation	[aussenisolation, gebäudeisolation, fassadenisolation, dachisolation, neuisolation, nachisolation, isolation, aussenisolationen, giebelisolation, zusatzisolation, innenisolation, gesamtisolation, isolationen, wärmeisolation, fassadenisolationen, fassadenaussenisolation, dachraumisolation, dachaussenisolation, bodenisolation, innenisolationsverkleidung, dachisolationen, aussensisolation, wandisolation, zwischenisolation, perimeterisolation, innenisolationen, vollisolation, weiterisolation, innenisolationsschale, ausenisolation, dachbodenisolation, gebäudeaussenisolation, originalisolation, isolationssparren, ueberdachisolation, aussenwandisolation, unterdachisolation, isolationspaneele, glasisolation, isolationsverputz, kellerdeckenisolation, aufdachisolation, fassenisolation, flachdachisolation, estrichisolation, assenisolation, fassadeisolation, aussenisolationmit, isolationsplatte, isolationn, neueraussenisolation, dachstockisolation, zwisxhenisolation, innenisolationesn, isolationspanellen, sparrenisolation, isolationsfenster, isolationsarbeiten, kellerisolation, isolationsschicht, fussbodenisolationen, fassadenhausisolationsverkleidung, zwischensparrenisolation, isolationfläche, isolationsputz, isolations, neuerisolation, deckenisolation, dachinnenisolation, estrichbodenisolation, inneisolation, isolationsplatten, ausseniisolation, zwichenisolation, mettalpanelaussenisolation, zwischenisolationen, mineralwollenisolation, aussenisolationsfassade, faddasenisolation, aussisolation, ausseisolation, fassadenisolationsverkleidung, aussenisolationund, kompaktisolation, isolationf, sockelisolation, mitaussenisolation, innenisolationa, isolationsglastüre, neueaussenisolation, innenisolationnen, isolationsverstärkung, innisolation, isolationb, minergiestandardisolation, sandwichpaneelisolation, wandkasettenisolation, isolationsverbesserung, teilisolation, isolationd, undisolation, wandisolationen, ausssenisolation, satteldachisolation, fassaenisolation, dchisolation, aufisolation, innemisolationen, dachisolationsverkleidung, hochleistungsisolation, dachnachisolation, raumisolation, steildachisolation, isolationserneuerung, mineralisolation, isolatione, fasssadenisolation, fassadisolation, spezialisolation]
isolierung	[wandisolierung, isolierung, fassadenisolierung, aussenisolierung, wärmeisolierung, nachisolierung, dachisolierung, neuisolierung, innenisolierung, dachinnenisolierung, isolierungverglasung, gebäudeisolierung, innisolierung, flachdachisolierung, dachsisolierung, wändeisolierung, inneisolierung, innenisolierungen, innnenisolierung, isolierungen, lukarneisolierung, deckenisolierung, gebäudeinnenisolierung, zwischenisolierung, fussbodenisolierung, teilisolierung, perimeterisolierung, gebäudehüllenisolierung, aufisolierung, ausisolierung, hausisolierung, isolierungsplatten, thermoisolierung]
dreifach	[dreifachverglasung, dreifachverglast, dreifachverglaste, dreifachverglasten, dreifachverglasungen, dreifachglas, dreifachisolierfenster]
verglasung	[isolierverglasung, dreifachverglasung, wärmedämmverglasung, photovoltaikverglasung, dreifachverglasungen, isolierverglasungen, isolierverglasungsfenster, wärmeschutzverglasung]

word part	*word list*
photovoltaik	[photovoltaik, photovoltaikanlage, photovoltaikmodulen, photovoltaikzellen, photovoltaikanlagen, photovoltaikmodule, photovoltaikanlaege, photovoltaikanage, photovoltaikplatten, photovoltaikpaneele, photovoltaikanalge, photovoltaikflächen, photovoltaikpaneelen, photovoltaikmodul, photovoltaikalnage, photovoltaikfeld, photovoltaikanlge, photovoltaiklanlage, photovoltaikinselanlage, photovoltaikanlange, aufbauphotovoltaikanlage, photovoltaikablage, photovoltaikanalage, photovoltaikpanels, grossphotovoltaikanlage, photovoltaikanlagenaufbau, photovoltaikanlgel, dachphotovoltaikanlage, photovoltaikaufbau, photovoltaikangaben, photovoltaikkollektoren, photovoltaikanlafe, photovoltaikgläser, photovoltaikanlageauf, photovoltaikelementen, photovoltaikdächer, photovoltaikanlag, photovoltaikanllage, photovoltaiksanlage, photovoltaikflachen, photovoltaikdach, photovoltaikanla, photovoltaikverglasung, indachphotovoltaikanlage, photovoltaikanlagn, photovoltaikananlage, photovoltaikpanel, photovoltaikanlanlage, photovoltaikeindeckung, photovoltaikfassade, photovoltaika, anbauphotovoltaikanlage, photovoltaikpanelen, fussbodenheizungphotovoltaikanlage, photovoltaikelemente, photovoltaikremise, photovoltaikfassaden, photovoltaikfasssaden, photovoltaikzaun, photovoltaikannlage, photovoltaiksolarmodule, photovoltaikanlagel, indachphotovoltaik, aufdachphotovoltaikanlage, gemeinschaftsphotovoltaikanlage, photovoltaikanlagearbeiten, photovoltaikaanlage, photovoltaikindachanlage, photovoltaiksolarsystemanlage, planungphotovoltaikanlage, pphotovoltaik, photovoltaikanlaga, photovoltaikalage, photovoltaikanlöage, photovoltaikanlasge, photovoltaikanölage, photovoltaikvolldach, photovoltaikanlae, fassadenphotovoltaikanlage, inndachphotovoltaikanlage, photovoltaikgeländer, photovoltaikange, photovoltaikianlage]
energetisch	[energetische, energetischer, energetischen, energetisch, energetischesanierung, energetischemassnahmen, energetisches, teilenergetische, eenergetische, energetischem]
dämung	[aussendämung, aussenwärmedämung, aussenfassadendämung, wärmedämung, zwischenwärmedämung, innenwärmedämung]
dammung	[wärmedammung, aussenwärmedammung, wämedammung, fassadendammung, aussendammung]
dmmung	[zwischenwärmeädmmung, aussenwärmeädmmung, innenwärmeädmmung, fassadendmmung, aussenwärmedmmungen, aussenwärmedmmung]
geotherm	[geothermischen, geothermischer, geothermische, geothermie, geothermiesonden, geothermieanlage, geotherm, geothermi]
thermie	[solarthermieanlage, solarthermie, geothermie, geothermiesonden, geothermieanlage, thermieanlage, solarthermieanlagen, solarthermiepaneelen, solarthermiermodule, géothermie, solathermie]
sonnenkol	[sonnenkollektoren, sonnenkollektorenanlage, sonnenkollektoranlage, sonnenkollektor, sonnenkollektors, dachsonnenkollektoranlage, dachsonnenkollektoren, sonnenkolletoren, sonnenkolleltoren, sonnenkolektoren, sonnenkollektroren, sonnenkollektorfeldes, sonnenkollektorenfelder, sonnenkoll, sonnenkollektorsystems, flachsonnenkollektoren, sonnenkollektoranlagen, sonnenkollektorflächen, dachsonnenkollektoranlagen, sonnenkollekoren, sonnenkollketoren, sonnenkollektorenanalge, sonnenkollekton, sonnenkollektorfeld, sonnenkollektorfelder, sonnenkolleketoren, sonnenkollektorenanlagen, sonnenkollektorananlage, sonnenkollektoen, sonnenkollektorten, sonnenkolloektorenanlage, sonnenkollektorenaufbau, sonnenkollektopren, sonnenkolldektoren, sonnenkollektorenwärmepumpe, sonnenkollektorenmontage, sonnenkollektoranage, sonnenkollekorpanels, sonnenkollktoren, sonnenkolektorenanlage, sonnenkollentoren, indachsonnenkollektoren, sonnenkollekorten, sonnenkollektoreanlage, sonnenkollektorenfeldes, sonnenkollektroen, aufbausonnenkollektoren, sonnenkollektoten, sonnenkollektren, flächensonnenkollektoren, sonnenkollektorern, sonnenkollekten, sonnenkollektorren, vonsonnenkollektoren, sonnenkollkektoren, sonnenkollektorer, sonnenkollektore, sonnenkollektorenelemente]
holzfeuerung	[holzfeuerung, holzfeuerungsanlage, stückholzfeuerung, hackholzfeuerung, holzfeuerungsofen, holzfeuerungs, pelletsholzfeuerung, holzfeuerungsnennleistung, kleinholzfeuerung, holzfeuerungsleistung, holzfeuerungen, stützholzfeuerung, holzfeuerungsraum]
holzheizung	[holzheizung, stückholzheizung, stückholzheizungsanlage, ersatzholzheizung, holzheizungsanlage, stückgutholzheizung, stzückholzheizung, holzheizungen, zusatzholzheizung, pelletholzheizung, stückholzheizungsraum, restholzheizung, stützholzheizung, sütckholzheizung, stückholzheizungseinbau, stckholzheizung, scheitholzheizung, stückhholzheizung, stückelholzheizung, sückholzheizung, schnitzelholzheizung, stüchholzheizung]
holzschnitzelheizung	[holzschnitzelheizung, holzschnitzelheizungsanlage, holzschnitzelheizungs, altholzschnitzelheizung, holzschnitzelheizungheizung, hackholzschnitzelheizung, holzschnitzelheizungsraum, holzschnitzelheizungen]
pelletheizung	[pelletheizungsanlage, pelletheizung, holzpelletheizung, pelletheizungen, stückholzpelletheizung]
komfortlüftung	[komfortlüftung, komfortlüftungsanlage, komfortlüftungssystem, komfortlüftungen, teilkomfortlüftung]
freecooling	[freecooling]
sonnenschutz	[sonnenschutz, sonnenschutzes, sonnenschutzverglasung, sonnenschutzstoren, sonnenschutzsystem, sonnenschutzanlagen, sonnenschutzglas, sonnenschutze, sonnenschutzelementen, sonnenschutzeinrichtungen, sonnenschutzmarkise, sonnenschutzlamellen, sonnenschutzvorrichtung, sonnenschutzdach, textilsonnenschutz, stoffsonnenschutz, sonnenschutzanlage, sonnenschutzbedachung, sonnenschutzrollos, sonnenschutzt, sonnenschutzmassnahmen, sonnenschutzmarkisen, sonnenschutzelemente, sonnenschutzsegel, sonnenschutzeinrichtung, sonnenschutzroste, sonnenschutzkonstruktionen, sonnenschutzfolien]
minergie	[minergie, minergiestandard, minergiestandart, minergiehaus, minergiebauweise, minergiezertifikat, minergiehäusern, minergiestand, minergiehäuser, minergiestandards, minergiegebäude, minergielabel, minergiewohnung, minergiestadard, minergieeinfamilienhaus, minergiebau, minergiestanard, minergiesanierung, minergiestandad, minergievorschriften, 8minergiestandard, minergiestaqndard, minergiehauses, minergiestandardhäusern, nminergiehaus, minergiestandardbauweise, minergiestandar, minergiep, minergies, minergiestsandard, minergiewerte, minergiewert, minergieneuhaus, minergiestandrad, minergiestandardisolation, minergiesstandard, minergiefenster, minergiehauss]
dachbegrü	[dachbegrünung, flachdachbegrünung, dachbegrüng, dachbegrünungen, dachbegrünnung]
fensterer	[dachfensterersatz, fensterersatz, fensterersetzung, fenstererneuerung, fenstererstaz, fenstererneuerungen, fensterersetzen, fensterersat, fenstererneuerund, fenstererstatz, teilfensterersatz, fenstereratz, dachflächenfensterersatz, einzelfensterersatz]
vip	[]
kaskadenlüftung	[]
verbundlüftung	[]
geocooling	[]

D.3. **Rules to Identify Energetic Restoration Activities.**

Three new classes of renovation activities were created that best matched the defined energetic restoration activities asked in the survey of TEP Energy: window replacement, energetic facade renovation, and energetic roof renovation (Jakob et al., 2021, pp. 33-36). The new classes were identified by using the text and the preexisting categories. A building application of the application type renovation was classified as window replacement if either the text contained a word of the word part *dreifach, verglasung* or *fensterer* (the complete word list of the word parts is given in Table D.3), or the building application was assigned to a preexisting category associated with window replacement, that is 300, 301, 302, 303, 304, 306 or 310 (the decoding is given in Table D.1). Analogously the rules for the two other classes (energetic facade renovation and energetic roof renovation) are given in Table D.4.

TABLE D.4. Classification Rules with Text and Preexisting Categories

Class	***Word part***	***Category***
Window replacement	dreifach, verglasung,fensterer	300, 301, 302, 303, 304, 306, 310
Energetic facade renovation	fassade	200, 202, 205, 207, 211, 214, 215
Energetic roof renovation	-	104, 109

E.1. **Evaluation of Energetic Restoration Classification Based on Preexisting Categories.**

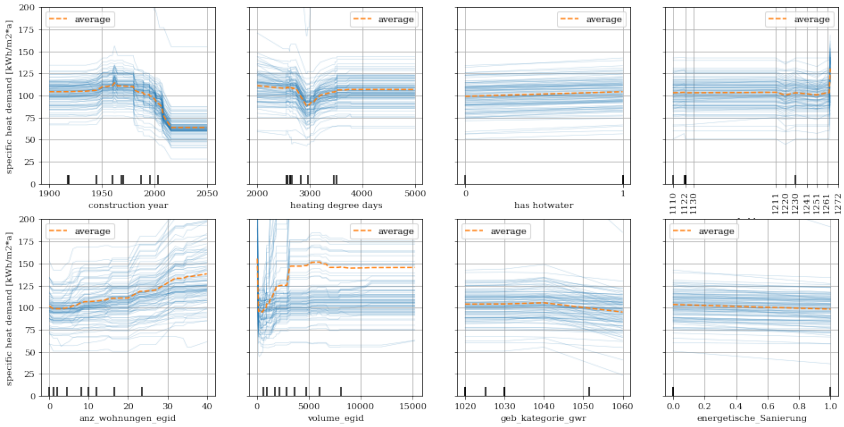


Figure E.1. Partial Dependency Plots of Heat Energy Demand Model

E.2. **Evaluation of Classification by Text.**

The classification based on the text data of the building applications is compared to the classification by given categories. Several classes can be compared, and we show here the comparison of the classes window replacement (Figure E.2), roof greening (Figure E.3), district heating (Figure E.4), geothermics (Figure E.5), comfort ventilation (Figure E.6) and minergie (Figure E.7).

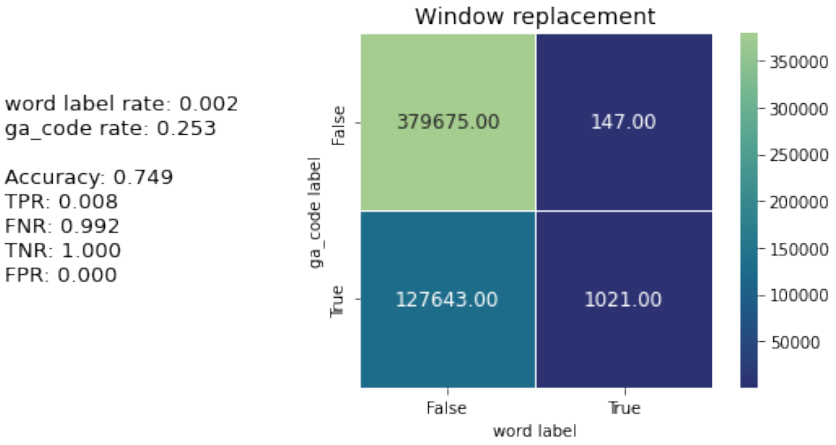


Figure E.2. Window Replacement: Comparison of Word and ga_code Label

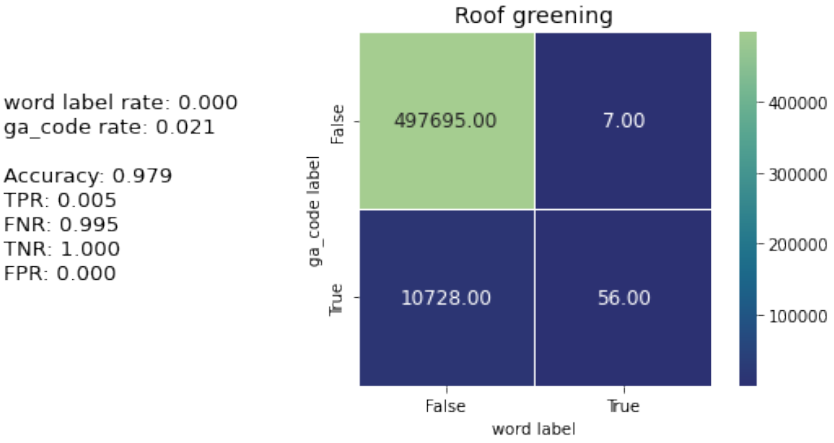


FIGURE E.3. Roof Greening: Comparison of Word and ga_code Label

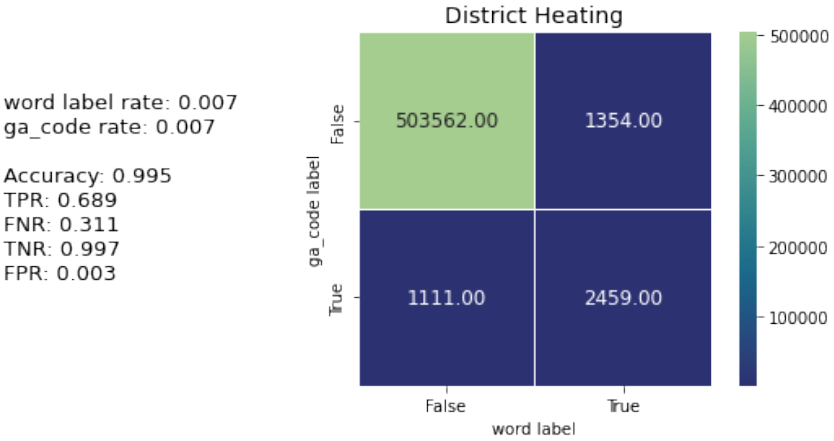


FIGURE E.4. District Heating: Comparison of Word and ga_code Label

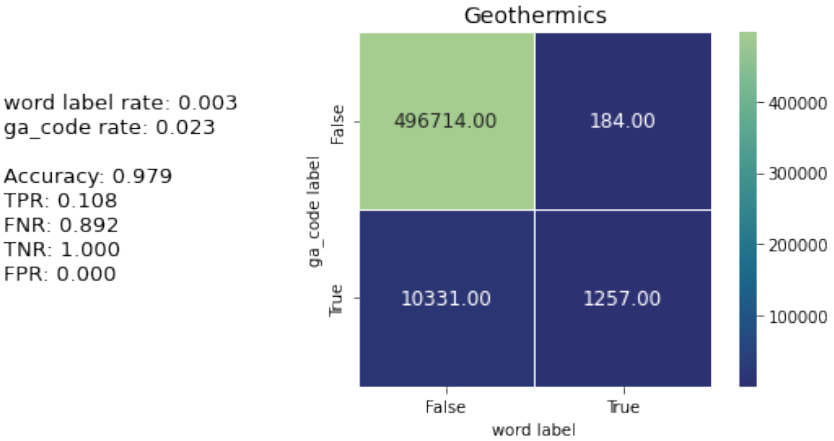


FIGURE E.5. Geothermics: Comparison of Word and ga_code Label

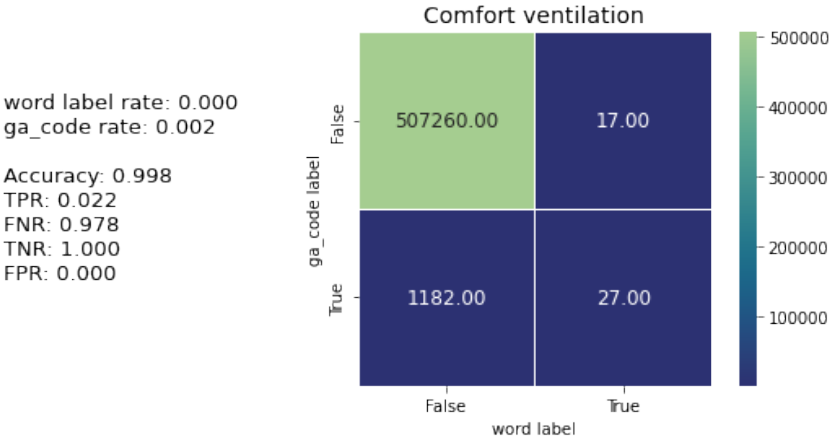


FIGURE E.6. Comfort Ventilation: Comparison of Word and ga_code Label

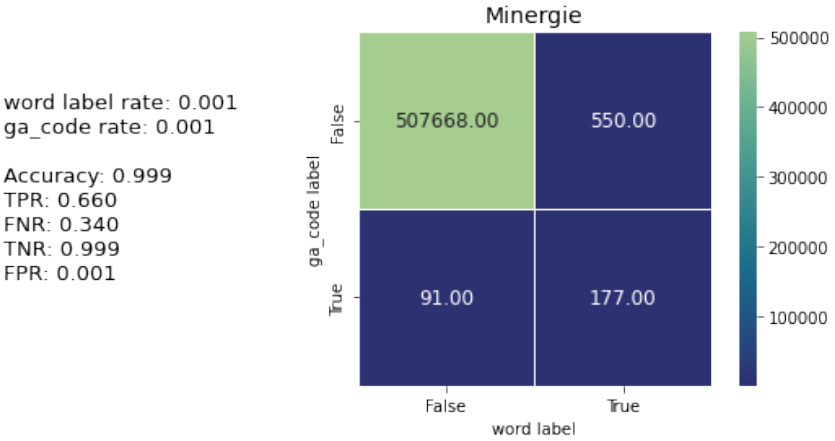


FIGURE E.7. Minergie: Comparison of Word and ga_code Label

E.3. **Feature Importance.**

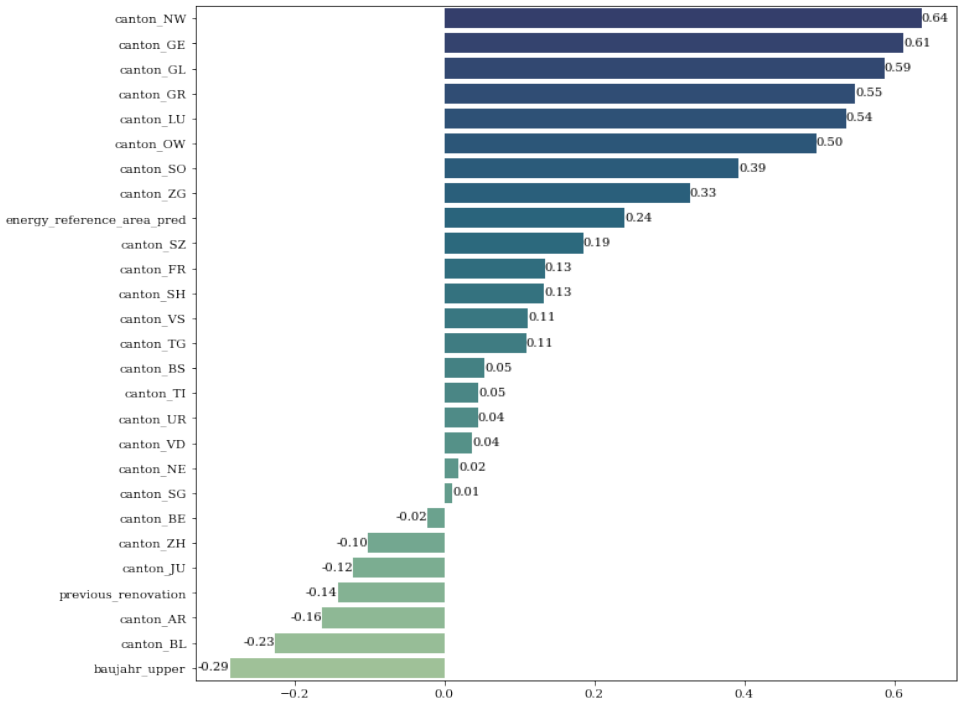


FIGURE E.8. Feature Importance of Logistic Regression Simple Model

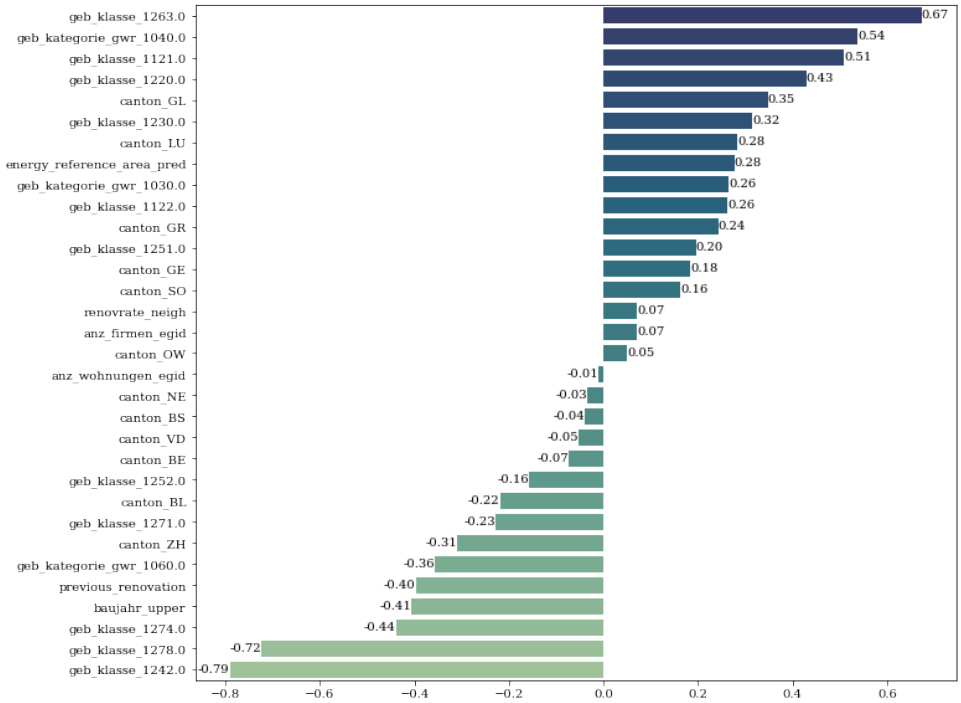


FIGURE E.9. Feature Importance of Logistic Regression Class-Weight Undersampling Model

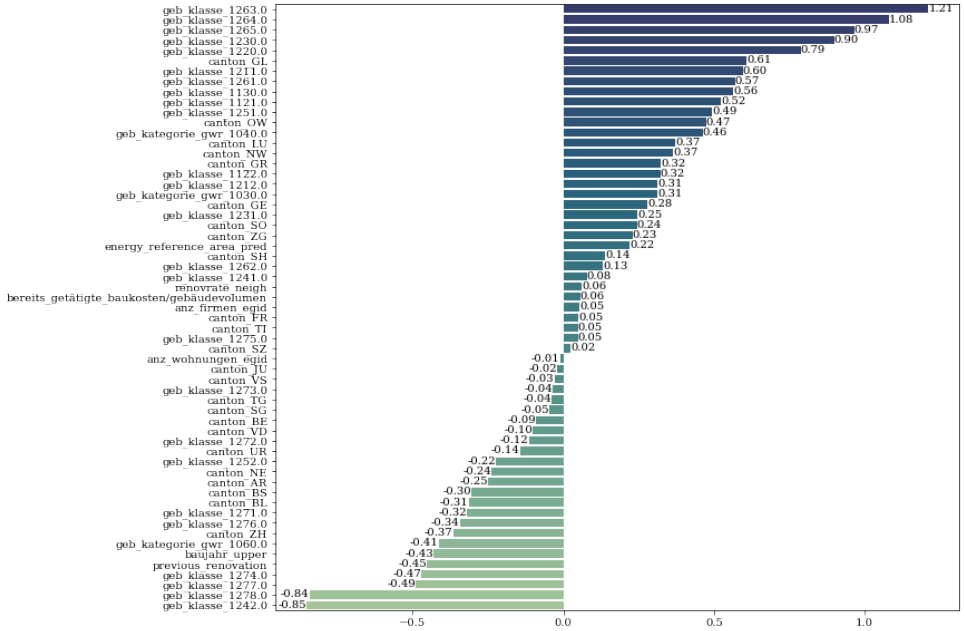


FIGURE E.10. Feature Importance of Logistic Regression Random Undersampling Model