ETH zürich

Institute for Dynamic Systems and Control

Martin Inauen

Non-intrusive occpuancy detection in smart buildings

A data-driven modelling approach

Semester Project

Institute for Dynamic Systems and Control
Swiss Federal Institute of Technology Zurich

Supervision

Michael Locher (EMPA)
Simon Muntwiler
Prof. Dr. Melanie Zeilinger

January 2023

IDSC-XX-YY-ZZ

IDSC

Abstract

Knowing the occupancy of a room is essential to improve space utilisation and energy optimization in buildings. Today it is still a big challenge to gather this data due to the lack of special-purpose sensors and the increasing importance of data privacy. In this semester project, a data-driven model was developed to predict the number of people in a room from different sensor inputs. The developed models are evaluated and compared on their performance. By using feature engineering on the different sensor measurements, it was possible to improve the model in bad-performing regions. Different strategies to extract more information from the CO_2 sensor are shown. Furthermore the dependency of the model prediction and the CO_2-measurement is investigated. The model was then evaluated on the data of other domains and future improvements of the model are discussed.

Keywords: occupancy detection, non-intrusive, smart buildings.

Contents

Chapter 1

Introduction

1.1 Motivation

Minimizing energy consumption has become increasingly important all over the world. The building sector is responsible for approximately 40% of the total energy consumption[1] and 30% of the greenhouse gas emissions[2]. In recent years, the application of smart buildings has become a trend. Having more information about the status of a building such as indoor air quality, energy consumption or air ventilation this data can processed and used to make certain decisions. As shown in [3] the HVAC energy consumption can be reduced by up to 23% by implementing an occupancy profile based control strategy. Therefore, it is important to extract occupant behaviour from the available sensors.

Human beings in indoor environments can be detected with the use of different sensors, such as passive infrared (PIR) sensors, video-cameras or device-free localization based on radio signals[4]. A lot of these techniques reveal more information than needed for occupancy detection and hence can be a risk for data privacy. To protect the privacy, it is crucial to just use non-intrusive sensor information to extract the occupant behaviour.

1.2 Goal

This project was initiated by an external partner of EMPA. The external partner developed a model to estimate the number of people from the CO_2-level and the ventilation airflow. This model is based on the mass balance equation of carbon dioxide. In non-dynamical systems (e.g. closed rooms) this model preforms relatively well, whereas in highly-dynamical systems with opened windows and doors this model does not perform well. Hence, the focus of this work was on improving the prediction by implementing a data-driven machine learning approach. This model can then be combined with the prediction of the model from the external partner to obtain a better model. The data-driven model should be compensating for the model mismatch of the model of the external partner. Ideally the model should be adaptable to other domnains.

Chapter 2

Related Work

Based on the goals for this thesis, relevant literature was consulted and existing solutions to the problems were studied in detail. As shown in different studies [5, 3], CO_2 concentration is a strong predictor variable for occupancy. Physically inspired models[5] try to predict occupancy based on the mass-balance equation of the CO_2 concentration. This models rely on estimating different parameters of the model, such as respiration rate of the human body or the air-volume of the room. The results of this study showed, that it is crucial to estimate or measure the state of a room (if the windows/doors are open or not). In another paper[6], a Bayesian Markov chain Monte Carlo approach for occupancy estimation with immeasurable ventilation is introduced, where it was shown that the prediction heavily relies on the estimate of the ventilation rates and the uncertainty in CO_2 measurements. An overview of different studies that have been applying machine learning models to predict the occupancy and window-opening behaviour is given in the literature study [3]. The results of the study showed that it is important to use feature engineering or to create additional CO_2-related predictor variables. For the window-opening behaviour the outdoor temperature, the indoor temperature and the wind speed are the most important predictor variables. It was also shown that artificial neural network models have a high prediction accuracy, whereby their transferability to other rooms is in general not as good as that of traditional ML models.

In order to construct a machine-learning model which combines the white-box model of the external partner with a ML model different approaches can be used. A study about grey-box modelling [7] showed that in general grey-box models can be classified into serial approaches and parallel approaches. Different techniques to combine the black-box model with the white-box model are presented. In some cases it is beneficial to only model the residual or error of the white-box model with the black-box model, whereas in other areas it is better construct a grey-box model, where the white-box model is just an input into the black-box model.

After consultation of the relevant literature it was seen that there has already been a lot of research conducted in the field of occupancy detection. In most of the studies either a black-box modelling approach or a white-box modelling approach is chosen. This showed that there is a demand to combine the both to achieve an overall better performance. The research also showed that there is no obvious existing solution which is predicting the number of people exactly and is still transferable. The lack of transferable ML models added to the motivation of this project to construct a grey-box model which could be more easily transfered to other domains.

Chapter 3

Problem Definition

3.1 Task Definition

The goal of this project was to develop a machine learning model for occupancy detection in smart buildings. Based on historical data a supervised model should be able to predict the number of people in a room. To do so, different sensor inputs such as CO_2-concentration, temperature, the ventilation flow or outdoor weather information can be used. This sensors should not violate the privacy of the individuals. All the data is labelled and hence the model can be verified on the testset of the data. The availability of labelled data allows for a supervised machine learning approach.

Another important aspect of the constructed model is its transferability to other domains. The model should be scalable and generalizable. Hence, the constructed model is also tested on data from other rooms to show whether the model is transferable to other domains.

3.2 Model Architecture

The decision of the model architecture is driven by the project requirements of the external partner. The final model must be able to combine the predictions of the white-box model with a black-box model. The focus of this semester project was, to develope the data-driven machine learning model (black-box model) as a part of the final model.

3.2.1 Grey-box Model

A grey-box model usually consists of a white-box model and a black-box model. The white-box model uses knowledge such as rules or theories to formulate a model which represents a physical phenomena. Comprehensive knowledge of the target system is needed and the modelling part requires simplifying the real-world system to obtain a mathematical formulation. The process of simplification involves making idealized assumptions, neglect interactions of different quantities or estimating parameters. In contrast to the white-box model, a black-box model does not require knowledge of the physical phenomena and hence can be used on unkown systems. The predictabiliy of a black-box model heavily relies on the data quality and the modeling algorithm. Black-box models make use of statistical approaches to reduce inaccuracies that can arise from incomplete knowledge. Only learning the physical relations from data the problem can arise that the model is highly specific on a given domain.

The approach of grey-box modelling or hybrid modelling combines the benefits of domain knowledge and empirical information. The resulting model can obey physical rules while minimizing the model mismatch of the white-box model with the black-box model. The black-box and white-box model can be combined either in serial or in parallel.

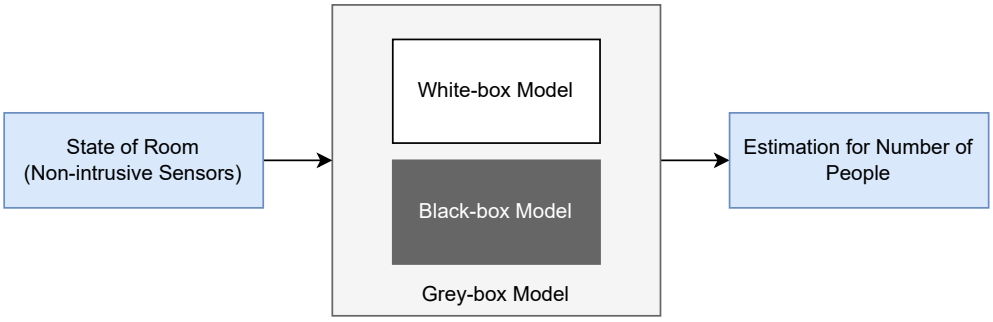


Figure 3.1: The combination of a white-box model with a black-box model results in grey-box model

3.2.2 Black-box Model

The developement of the black-box model was the focus during this semester project. In order to achieve the best possible results, different machine learning models were tested and evaluated. Since the output variable of the model is a discrete value, either a classification or a regression model can be used. A regression still is the more favorable because it could be that a class is missing in the training data and thus a regressor would be able to capture this correlation. By using a regression the adaption to other rooms is more likely to work as with classification.

Random Forest

Random Forest is an ensemble learning method, where a multitude of decision trees is constructed at training. The final prediction of the model in regression tasks usually is the average of the predictions of the individual trees. This model uses the technique of bootstrap aggregating, or short bagging, in order to decrease the variance of the model without increasing the bias. In bagging, the model repeatedly selects a random sample with replacement and fits trees to this samples. This bagging technique can also be used to choose a random set of feature for each tree and fit individual trees on each subset. While the random forest reduces overfitting in decision trees, it requires much computational power to build numerous trees.

Gradient-Boosted Decision Trees

This model works by building simple (weak) prediction models sequentially, where each model tries to predict the error left over by the previous model. This model allows to optimize with respect to an arbitrary differentiable loss function. The disadvantage of this model is the lack of intelligibility and interpretability, since it is hard to tell why the model has made a certain decision. During this project *XGBoost*[8] was used, which is an implementation of gradient-boosted decision trees. *XGBoost* stands for "Extreme Gradient Boosting", which obtains the same results as gradient boosting but is able to do parallel learning and hence is more efficient.

Chapter 4

Experimental Evaluation

As the biggest part of the semester project was the developement of the supervised machine learning model, the experimental evaluation makes up for the biggest part. This section is structured into a methodology part, where the whole experimental setup is explained and the different steps to develop the model are presented. In the result section the constructed model is evaluated and compared to the white-box model. After evaluating the model on one room, the model is tested on other domains and results are shown.

4.1 Methodology

The general methodology of building a machine-learning model can be split up into the different steps seen in figure 4.1. After the data collection all the data is pre-processed. The last three steps are then conducted iteratively, where different feature engineering and modelling approaches are tested.

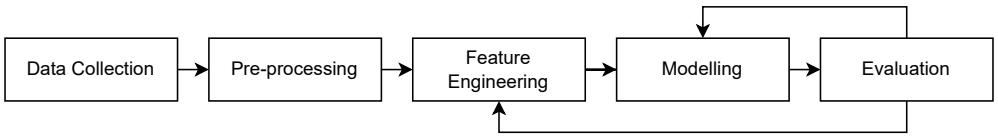
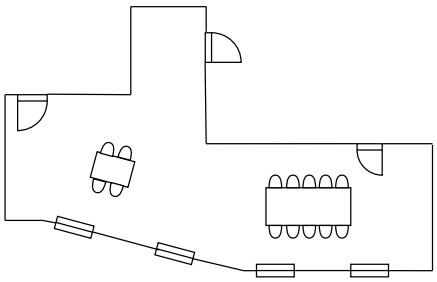


Figure 4.1: The different steps to build a black-box model

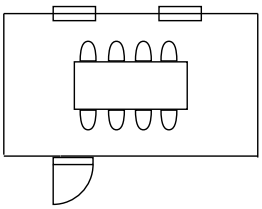
4.1.1 Experimental Setup

The infrastructure for the experimental setup is located in Dübendorf, Zürich. NEST[1] is a modular research and innovation building of EMPA and Eawag. In this facility, different technologies, materials and systems are tested. Therefore the rooms used as testbeds are constructed of different materials and hence do differ in their characteristics (e.g. thermal absorption, CO_2 leakage, ...). NEST is fully equipped with lots of different sensors and hence the focus of the data acquisiton was to get the ground-truth of people present in a room. This was done by placing a people-counting system (camera) [2] above each door. Since this cameras were prone to errors, the data had to be manually labelled again ande checked for correctness. Also for all the sensors it had to be ensured that they work well and that the data is complete.

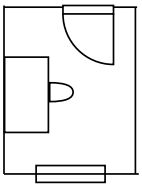
[1] `https://www.empa.ch/de/web/nest/`
[2] `https://www.xovis.com/technology/sensor`



(a) Room A, a big conference room



(b) Room B, small conference room



(c) Room C, small office

Figure 4.2: The three different testbeds used for the experiments

In the experiments the data from three different testbeds as seen in figure 4.2 was collected. This testbeds differ significantly in size and occupation and hence build a great foundation to compare the model on different domains. The air sensors present in the different testbeds do differ in quantity as seen in table 4.1. Besides the measurements of the air the data of opened windows and opened doors was collected. Also the outdoor weather was sensed with different sensors, such as temperature, humidity and CO_2-concentration.

All the data was collected and labelled for the duration of 4 months. The data is sampled with a rate of $1/\,\mathrm{min}$.

Sensor Type	Room A	Room B	Room C
Temperature	10	1	5
CO_2	2	1	1
Humidity	10	1	1
VOC	4	2	1
Ventilation Flow	2	1	1

Table 4.1: Type and quantity of air-data sensors available in the different testbeds

4.1.2 Pre-processing

After the data collection the data had to be pre-processed. In the pre-processing step the data is resampled and homogenized. This step is crucial to achieve good results later. For the ground-truth of the occupancy some validation checks were made to make sure the model only receives data of good quality. If the occupancy was negative, the recordings were double-checked and the data was corrected. Furthermore the data was checked for gaps or double values and sensor measurements were observed to detect malfunction. If a sensor was not working correctly, the gaps in the series where identified and the data was interpolated. If the gap was bigger than 60 minutes, the data was deleted.

In order to obtain a model which is transferable to other rooms, the different sensors of the same type had to be aggregated to a single feature. This was done by taking the median of all the different sensors of the same type, because the median is less sensitive to outliers than taking the average.

4.1.3 Feature Engineering

As a first step of feature engineering a black-box model was trained with all the sensors as a feature. After the model has been trained, the features can be analyzed by their feature importance or the shapley value of the feature. The shapley value is a game-theoretic approach and tells us how much

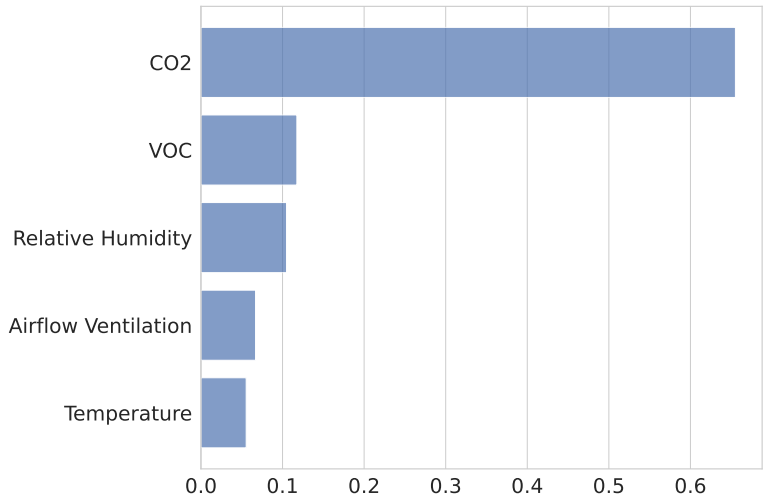


Figure 4.3: Feature importance of the aggregated features

each feature contributed to a given output.The evaluation of the feature importance can be seen in figure 4.3, the CO_2-concentration is by far the most important feature for the model.

Cyclical Features

Due to the cyclical nature of time-series data it can be helpful to construct a feature, which teaches the model the cyclical behaviour of the data. By encoding a day as a sine and cosine function from 0 to 2π the model will hopefully catch up the relation of the day as a repeating pattern.

In another feature, the weekdays from Monday to Friday were encoded with one, whereas the weekend was encoded with zero. The reason in doing so was that the likelihood of people present in the building is much lower at the weekends. In the same way the holidays of Switzerland were encoded and added as a feature.

Absolute Humidity

The humidity sensor of the testbed measures relative humidity. In favor of making this feature more interpretable, the relative humidity was converted to an absolute humidity with the use of the ideal gas law. A person emits approximately 1.2 liters per day, depending on the activity[9]. By using the absolute humidity, this finding can be used to hopefully construct a better feature than by just using the relative humidity.

To get more out of the absolute humidity feature, a Seasonal-Trend decomposition using LOESS (STL) of the original series was done as can be seen in figure 4.4. STL is a method to decompose time series into a seasonal, a trend and a residual component. LOESS stands for locally estimated scatterplot smoothing and is a generalization of the moving average and polynomial regression. The trend and seasonality window size have to be specified by the user. In the scenario of this project, the goal was to delete the long-term trend and the daily fluctuations from the absolute humidity to only get the remainder, which should reflect the disturbance coming from the occupants. The parameters of the STL decomposition were optimized to achieve the highest possible correlation between the occupancy and the residual of the STL.

Adding the residual as a feature to the model, it was seen that the importance of the feature is still quite low in comparison with the CO_2-concentration. This led to the conclusion that further feature engineering has to be done on the CO_2 feature.

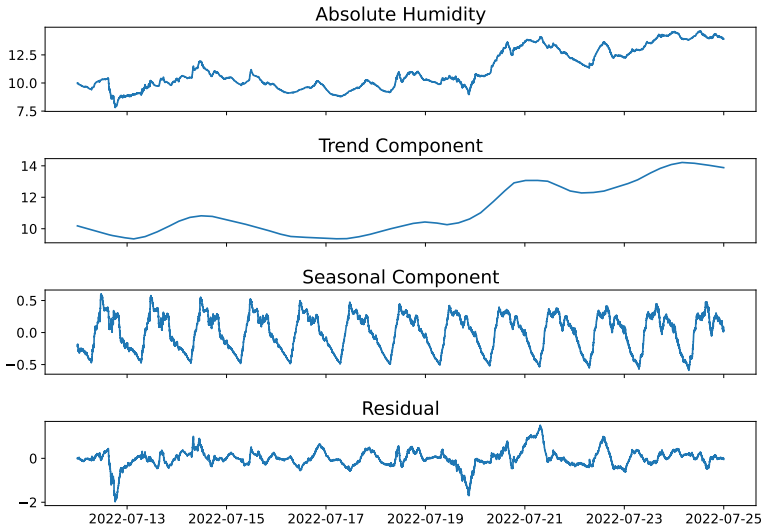


Figure 4.4: STL decomposition of the absolute humidity feature

CO_2

The CO_2-concentration was seen to be the best predictor variable for the prediction. This resulted in extensive feature engineering with this value, where different features derived from the CO_2-concentration were added to the feature space. When the human body is emitting CO_2, there is always a lag present until the sensor measures this change. By adding lagged versions of the CO_2-concentration to the feature space, this information can be incorporated into the model. In this project the lags of up to 5min were added as features. Furthermore the derivative of the CO_2-concentration was added to incorporate the dynamics. A smoothed version of the derivative was developed with the use of a Savitzky-Golay-Filter, which is basically a polynomial regression over a user-defined window and thus get a smoothed value for every point. This filter can also be combined with a derivation. This smoothed gradient was combined with a window-size of 30min to achieve a feature, which correlates well with the people present in the room as seen in figure 4.5

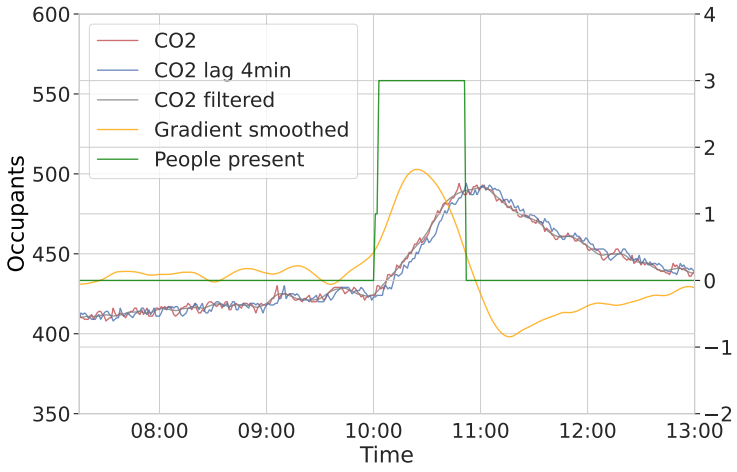


Figure 4.5: Different CO_2 feature used in model and people-present

4.1.4 Training and Evaluation strategy

First, the data was split in a train and test set with proportion 5:1, whereby the test set was always at the end part of the time series. After this, a randomized search combined with a cross validation was performed to do a hyperparameter optimization for the model. Because there exists a temporal dependency between the observations, a normal cross validation is not possible. Therefore, a time-series cross validation step was introduced in order maintain the temporal order of the data. The cross validation prevents the model from overfitting and is a more robust method to evaluate the performance. In the time-series cross validation, the training set was split up into a validation set and a training set as before, where the validation set was again the last part of the training set. The model was then trained on this set and evaluated on the validation set, while the training set was enlarged in size. This process is then repeated for k splits as seen in figure 4.6. For each split, the model is fit and evaluated and the evaluation metric of the model is calculated. For a given set of parameters, the metric is finally calculated as the average over all the splits. This procedure is repeated for different random sampled hyperparameter combinations to find the best parameter set.

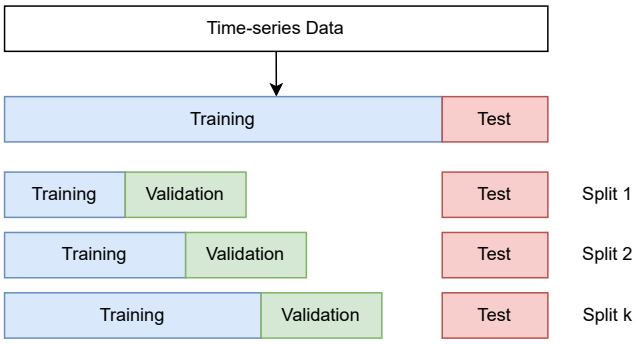


Figure 4.6: Time-series cross validation

In the final evaluation different metric were evaluated and compared over different iterations of the model developement. The most important ones were the root mean square error (RMSE), mean absolute error (MAE) and the R^2-score defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{x_i - y_i}{\sigma_i}\right)^2}$$

$$MAE = \sum_i |x_i - y_i|$$

$$R^2 = 1 - \frac{\sum_i (x_i - y_i)^2}{\sum_i (x_i - \bar{x})^2}$$

where y are the predicted values and x the observed values, with $\bar{x}$ being the mean of the observed values and σ being the standard deviation.

4.1.5 Model comparison

On both models, the Random Forest and the XGBoost model, a randomsearch was performed to find the best hyperparameters for each model. Both models were trained on the same dataset with equal features and an evaluation was made. Besides having a much lower computation time, the XGBoost model outperformed the Randomforest as seen in table 4.4

Model	R2	MAE	RMSE
Random Forest	0.32	0.93	2.66
XGBoost	0.37	0.90	2.54

Table 4.2: Comparison of the tested machine-learning models

Furthermore the shorter training duration of the XGBoost model allowed faster developement iterations. That and the better performance lead to the decision to continue the developement only on the XGBoost model.

4.2 Results

After lots of iterations of feature engineering to training and evaluating the model a promising model was obtained. The performance of the model is investigated in detail in the following section. Furthermore the adaption to other domains is tested and evaluated.

4.2.1 Final Model

A big improvement of the model was achieved by changing the learning task from a regression with squared loss to a poisson regression. The poisson regression is mainly used for count data and outputs the mean of the poisson distribution. This brings the benefit of avoiding the prediciton of negative values. The poisson regression can be used, when the target variable follows the poisson distribution. The poisson distribution is characterized with mean and variance being equal.

The performance of the final model is compared with the baseline model and the white-box model of the external partner. For the evaluation metrics mentioned in section 4.1.4 the results state as follows:

Model	R2	MAE	RMSE
White-box Model	0.26	1.48	2.89
Baseline Model	0.31	1.10	2.67
Final Model	0.39	0.81	2.52

Table 4.3: Evaluation metrics of White-box model, Baseline model and Final model

In the beginning of the developement a big problem was the relaxation behaviour of the model. When people left the room, The prediction just slowly went down. This problem was faced with advanced CO_2 feature engineering as explained in section 4.1.3. The decays of the CO_2-concentration after an occupancy event were detected and encoded such that it reflected that the people left the room at the beginning of the decay. Furthermore the model prediction was sometimes changing rapidly, which is not the case in the target variable. This big changes in prediction was improved by adding different versions of the CO_2-concentration, such as smoothed versions and a smoothed gradient. This improvements can clearly be seen in figure 4.7. The final model is reacting faster after people left the room than the baseline model. Further the prediction of the final model is less sensitive, meaning that the fluctuation of the final model is smaller compared to the baseline model.

Another problem of the model was the tendency to predict zero occupants, since most of the time the room is empty. Looking at the dataset, it can be seen that the data is highly unbalanced, since the zero-class is making up for around 81% of the total dataset. Rounding the prediction of the model to discrete values, the predictions can be analyzed with classification metrics. Evaluating the accuracy of the model, meaning the number of correct predictions divided by the total number of predictions, a accuracy of 0.78 was achieved. A model which only predicts zero values all the time would achieve a value of 0.81 on the same dataset. This is why the balanced accuracy is evaluated

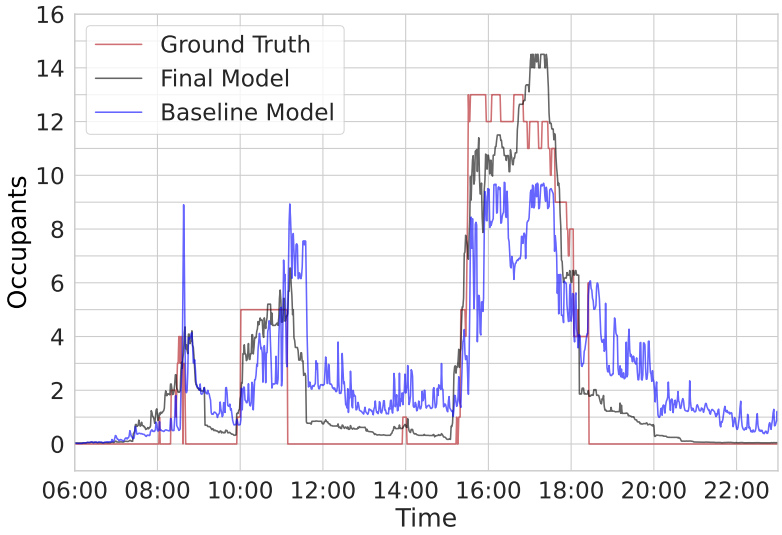


Figure 4.7: Comparison of initial baseline model and final model

as well, which takes into account that the data is unbalanced. This leads to a value of 0.09, where for the zero-predicting model the balanced accuracy would be 0.03.

Using the final model to predict only the states "occupied" or "not occupied" in a room leads to an accuracy of 0.89. This is achieved by mapping the regressor outputs > 0.5 to an occupied state. The balanced accuracy leads to a value of 0.82.

After analyzing the evaluation metrics of the model, further investigation was conducted on the exact distribution of the predicted values. In figure 4.8 the continuous prediction distribution for the different classes are visualized. In the range of 0 to 13 occupants it can be seen that the model is able to catch a relation. The distribution of the prediction is getting bigger with a bigger amount of people present. From the plot it can be obtained that the model tends to underestimate the number of people present, since it is beneficial for the model performance.

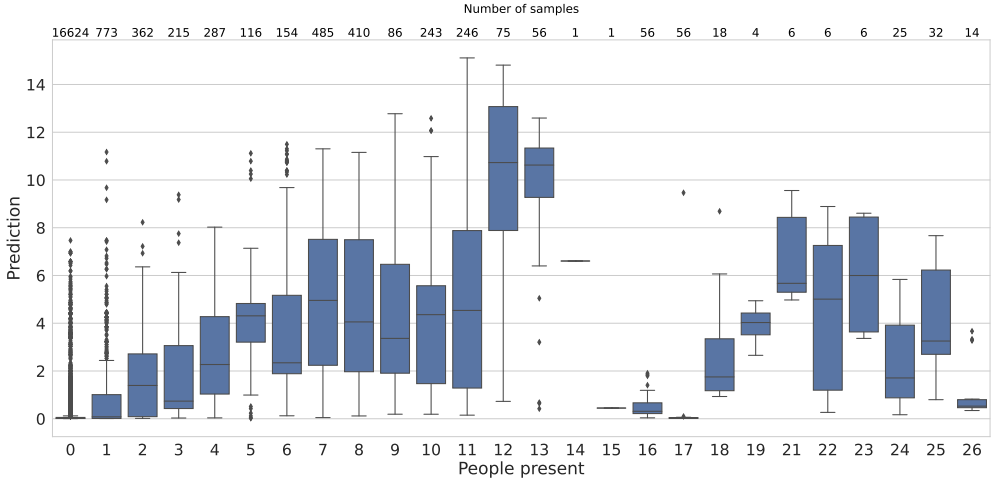


Figure 4.8: Prediction distribution of different classes and their sample sizes.

On the top of the plot in figure 4.8 the number of samples of each class can be found. It can be seen that there exists a negative correlation between values of occupants and the number of samples. From 13 and more occupants there is few data available, which is also one reason why the model is not able to perform well in these regions. Apart from this, the different samples in the training and test data were compared. It was found that in the training data a maximum of 20 people were present in the room, while in the test data a value of up to 26 was observed.

Looking at the predicted values in detail, it can be seen that while in some areas a good prediction is achieved, in others the prediction is still fluctuating fast. Because the evaluation of the importance of the feature showed a high dependency from the model of the CO_2-concentration, it can be assumed that this comes from the CO_2 signal. In figure 4.9, in the upper half the CO_2 signal is slowly building up, then it is fluctuating and afterwards decaying. Tis most probably comes from open doors or people moving around which leads to the mass of the air being moved around. In the lower plot this behaviour looks much more distorted over the whole occupancy event. From the collected data it can be seen that a window was open from 11:00 to 15:50 in the lower plot, while in the upper part the windows were closed. The fluctuations in the end peak probably come from an open door and thus a more active circulation of air.

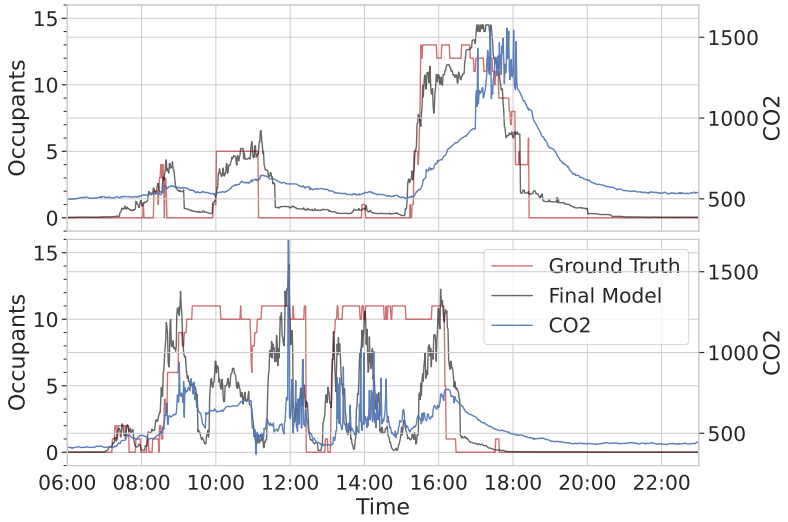


Figure 4.9: In the upper plot the CO_2-concentration and model prediction in a closed room, in the lower plot the CO_2-concentration and model prediction with open window.

As the model prediction and the CO2-measurements are highly correlated, a further analysis of the CO2-measurements was made. In figure 4.10 the same pattern as in figure 4.8 can be found. As can be seen from the median, the Number of people and the CO_2-level correlates relatively well up until 7-8 people, where we see a slight drop in the CO_2-level. We can observe a spike at 12 people, whereas from there the CO_2-level does not visibly correlate withe the people present anymore. This corresponds to the observed behaviour found in figure 4.8, where it was seen that the prediction gets worse for a higher number of people present.

In order to explain the behaviour found in figure 4.9 the predictions were analyzed depending on whether the room is in an open or a closed state. The room was classified as an open state, if one of the windows was open and as closed if all the windows were closed. The opening of the door was not part of the classification, as every time at least at the beginning and end of a presence event the door will be open. This classification showed that for the smaller predictions, the CO_2-measurements have a quite similar relation for the open versus the closed system. However, for increasing predictions the two curves of the median tend to become distant. For the closed system the CO_2-concentration reaches higher values as in the open system, but for the closed room the sample size is getting smaller and smaller for higher predicitons. For example for 9 people present,

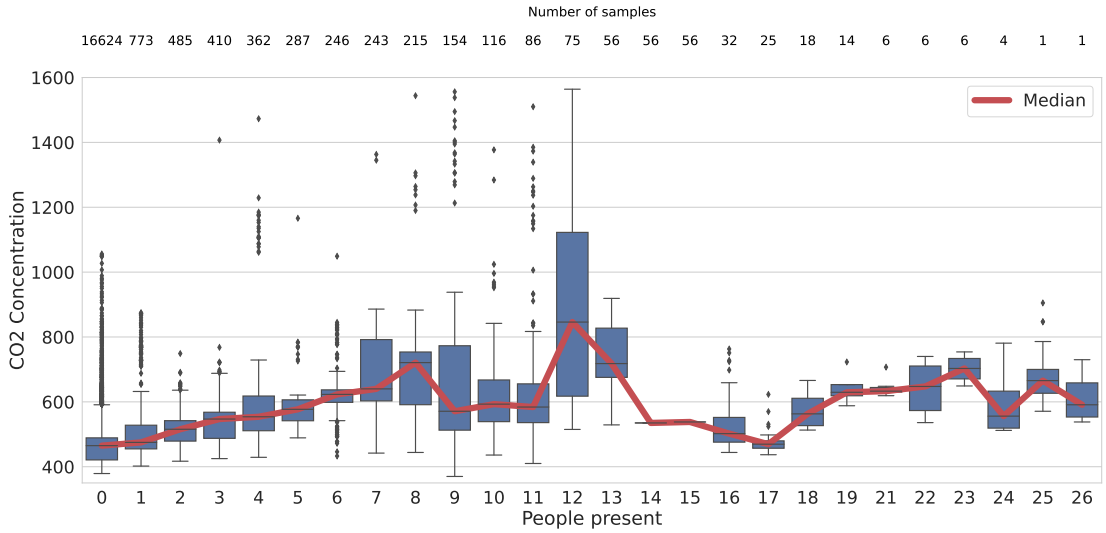


Figure 4.10: The distribution of the CO_2-measurements with respect to the number of people present.

the samples of the closed room were 32, whereas for the open room 184. With more people present the likelihood that one opens a window is much higher than with few people present. For the open system, there is a small increase in CO_2-measurement for higher number of people, but the value normally not exceeds 800ppm.

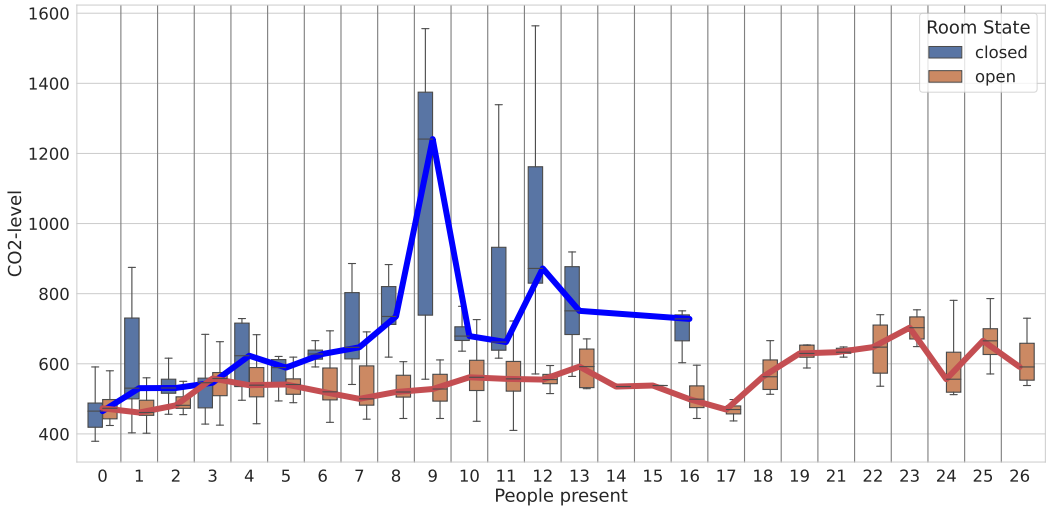


Figure 4.11: The absolute error of the prediction versus the people present in closed and open rooms.

4.2.2 Adaption to other Domains

The obtained model was developed and evaluated mainly on one room. To see whether the model is able to predict the number of people on other rooms as well, the model was tested and evaluated on the different domains mentioned in figure 4.2. To do so the available data of all rooms was collected and compared. Since there were different sensors present in the testbeds, the intersection

set of all the sensor types had to serve as feature space. From this the obtained features form the previous section were constructed. Compared to the final model described in the steps before, the ventilation and VOC information was not anymore part of the model. From this feature space, the hyperparameters were again optimized on the data of room A to obtain model A. The same principle was conducted to obtain models B and C, from the data of room B and room C respectively. This three models should then predict the output of the testdata of each room. The outcomes of this evaluation can be seen in table 4.4.

	Model A			Model B			Model C		
	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE
Room A	**0.36**	**0.82**	**2.58**	0.22	1.01	2.85	-0.09	1.13	3.38
Room B	0.57	0.34	1.01	**0.62**	**0.27**	**0.95**	-0.16	0.43	1.55
Room C	-196.12	2.11	3.90	-128.68	1.62	3.16	**0.52**	**0.06**	**0.19**

Table 4.4: Performance of Models trained on and evaluated on different Datasets

From this evaluation it can be seen, that the model A works relatively well on the data of room B and vice versa. However, for room C the predictions are far from truth, because the CO_2-concentration still lies in the same range but the people present do differ significantly from room C to room A. In room C, the maximum people present is just one person, while in room A the maximum is at 26 people. The prediction of model A on room C is therefore in another scale. A scaling approach of the prediction with respect to the maximum number of people present in a room was tried out but not found to be sufficient. It seems, that also the dynamics do differ from room A to room C and not only the scaling of the prediction. Model B works more or less on the data of room A, this can be explained that most of the time the people present are in the same range in room A and room B. In room A there are some extreme events with large amount of people present (>15), but those events only rarely occur. The model trained from the data in room C proved to be completely off for the other rooms, which corresponds to the finding that the data of room C does not work well on Model A and model B respectively.

As a pure scaling of the prediction output with respect to the maximum room occupation was not working well, a different approach was tested out. In figure 4.12, the model optimized and trained on room A was supplied with different amounts of training data of room C. It can be seen that even with only one day of training data available, a promising model can be obtained. As can be seen as a reference, the model C is the model with full parameter optimization on room C. From this we can see that the model not needs a lot of training data to be fit to a new domain.

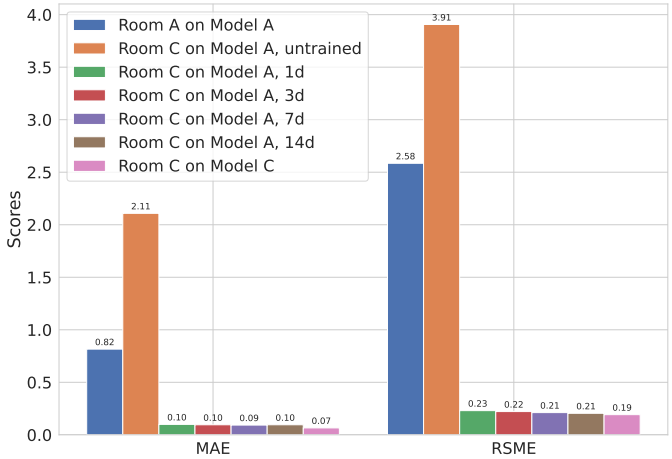


Figure 4.12: MAE and RMSE of Model A on data of other domain, fitted to different durations of data from new domain

Chapter 5

Conclusion and Future Work

A deeper understanding pf the occupancy pattern proves to be useful in many different areas, such as energy consumption reduction, room utilisation and security. To extract the occupancy pattern from non-intrusive sensor data it is especially important to protect the privacy of individuals. The model developed during the duration of this project can extract this information within the accuracies and error metrics shown in section 4.2. Specialized feature engineering proved to be important in order to extract more information out the raw CO_2 measurements. It was also seen that the CO_2-concentration shows by far the highest correlation to the number of people present from all the different air measures.

The prediction of the model gets less accurate with increasing number of people present. This behaviour mainly stems form the fact that the system is more dynamic with more people present. Knowing that the system behaves like this, the occupancy estimation could be split up into bins of increasing size. For most applications in real life it would be sufficient to have an estimate wheter there are 3-5 or 15-20 people present in the room, whereby the exact number is not that important. For an improvement of the model, the number of airshifts of a room could be incorporated into the model or a estimation of the state (open or closed) could be implemented. In doing so, the model would hopefully also better predict higher numbers of people present, because the model can distinguish betwween open and closed room state. In addition to the classification of open and closed systems it would be helpful to implement an estimation of the confidence of the output. In doing so, the occupancy pattern classified in regions where the model is confident.

Furthermore it was seen that the domain adaption works on rooms of similar size and with similar dynamics, but fails in rooms a different amount of people present. Still it can be said that the feature engineering strategies discussed can also be used on other rooms. Only supplying the model with a small amount of training data already leads to big improvements in the MAE and RMSE. For future models different domain adaptation strategies could be conducted. As seen from the evaluation it would already be sufficient to follow a semi-supervised approach, where the amount of required training data could be reduced to a minimum.

Bibliography

[1] G. David and B. Lafont, "Energy efficiency in buildings business realities and opportunities," *World Business Council for Sustainable Development) Go to reference in article*, 2008.

[2] Y. Geng, W. Ji, Z. Wang, B. Lin, and Y. Zhu, "A review of operating performance in green buildings: Energy use, indoor environmental quality and occupant satisfaction," *Energy and Buildings*, vol. 183, pp. 500–514, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778818315378

[3] X. Dai, J. Liu, and X. Zhang, "A review of studies applying machine learning models to predict occupancy and window-opening behaviours in smart buildings," *Energy and Buildings*, vol. 223, p. 110159, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778820303017

[4] P. Vance, G. Prasad, J. Harkin, and K. Curran, "Analysis of device-free localisation (dfl) techniques for indoor environments," in *IET Irish Signals and Systems Conference (ISSC 2010)*, 2010, pp. 76–81.

[5] D. Calì, P. Matthes, K. Huchtemann, R. Streblow, and D. Müller, "Co2 based occupancy detection algorithm: Experimental analysis and validation for office and residential buildings," *Building and Environment*, vol. 86, pp. 39–49, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360132314004223

[6] H. Rahman and H. Han, "Estimation of occupancy in a naturally ventilated room using bayesian method based on co2 concentration," *International Journal of Mechanical Systems Engineering*, vol. 3, 11 2017.

[7] Z. Yang, D. Eddy, S. Krishnamurty, I. Grosse, P. Denno, Y. Lu, and P. Witherell, "Investigating grey-box modeling for predictive analytics in smart manufacturing," 08 2017, p. V02BT03A024.

[8] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, aug 2016. [Online]. Available: https://doi.org/10.1145%2F2939672.2939785

[9] T. Lu, X. Lu, and M. Viljanen, "Moisture and estimation of indoor moisture generation rate," in *Chemistry, Emission Control, Radioactive Pollution and Indoor Air Quality*, N. Mazzeo, Ed. Rijeka: IntechOpen, 2011, ch. 18. [Online]. Available: https://doi.org/10.5772/17014

ETH zürich

Institute for Dynamic Systems and Control

Prof. Dr. R. D'Andrea, Prof. Dr. E. Frazzoli, Prof. Dr. Lino Guzzella, Prof. Dr. C. Onder, Prof. Dr. M. Zeilinger

Title of work:

Non-intrusive occpuancy detection in smart buildings

A data-driven modelling approach

Thesis type and date:

Semester Project, January 2023

Supervision:

Michael Locher (EMPA)
Simon Muntwiler
Prof. Dr. Melanie Zeilinger

Student:

Name:	Martin Inauen
E-mail:	inauenma@student.ethz.ch
Legi-Nr.:	18-927-830
Semester:	HS 2022

Statement regarding plagiarism:

By signing this statement, I affirm that I have read and signed the Declaration of Originality, independently produced this paper, and adhered to the general practice of source citation in this subject-area.

Declaration of Originality:

`https://www.ethz.ch/content/dam/ethz/main/education/rechtliches-abschluesse/leistungskontrollen/declaration-originality.pdf`

Zurich, 14.1.2023: