Room Occupancy Detection

Charles Stuppard, Mergen Narangerel

12/4/2021

### Introduction

With the increase in popularity of smart homes and smart technology, smart sensors are being used more often. One such sensor, used for binary classification, is a room occupancy sensor. Such sensors operate on different measurements such as light illumination, temperature, and energy (electricity) output. In general, these occupancy sensors use a collection of sensory information to detect the presence of a person based on the values of each variable. Industrial applications of such sensors range from home security, energy savings and automation. To handle the automated prediction and classification of a person’s presence on a larger scale, supervised machine learning methods, such as decision trees, can be applied. These methods include logistic regression, Support Vector Machines (SVM), and Extreme Gradient Boosting, often referred to as “xgBoost”. These methods vary in approach and each contains their own set of advantages and drawbacks.

### Objective

The objective of this study is to comparatively apply Support Vector Machines and Extreme Gradient Boosting. Using the Occupancy Detection dataset, each method will be applied to accurately detect the presence of a person given updated information from the sensor variables in the test data. Although there have been previous comparison studies of SVM and xgBoost, this study applies these supervised classification methods on a dataset with an aggressively strong predictor, Light. This study further investigates the response when Light is removed and the remaining predictors (humidity, humidity ratio, CO2, and temperature) are modeled. The final outcome is expected to accomplish the following: present clearly defined differences between both classification algorithms and the logistic regression baseline; produce valid and accurate classification models using each method; and reinforce learning objectives regarding A.I. classification methods.

### Literary Review

There are several methods commonly used for machine learning classification problems; this study focuses on identifying a preferred machine learning classification method between support vector machines and extreme gradient boosting on the Occupancy Detection dataset. Generally, classification problems involve a qualitative, or categorical, response (ISLR). For this dataset, occupancy is determined by whether or not a person is detected in the room, based on the data provided by the sensors.

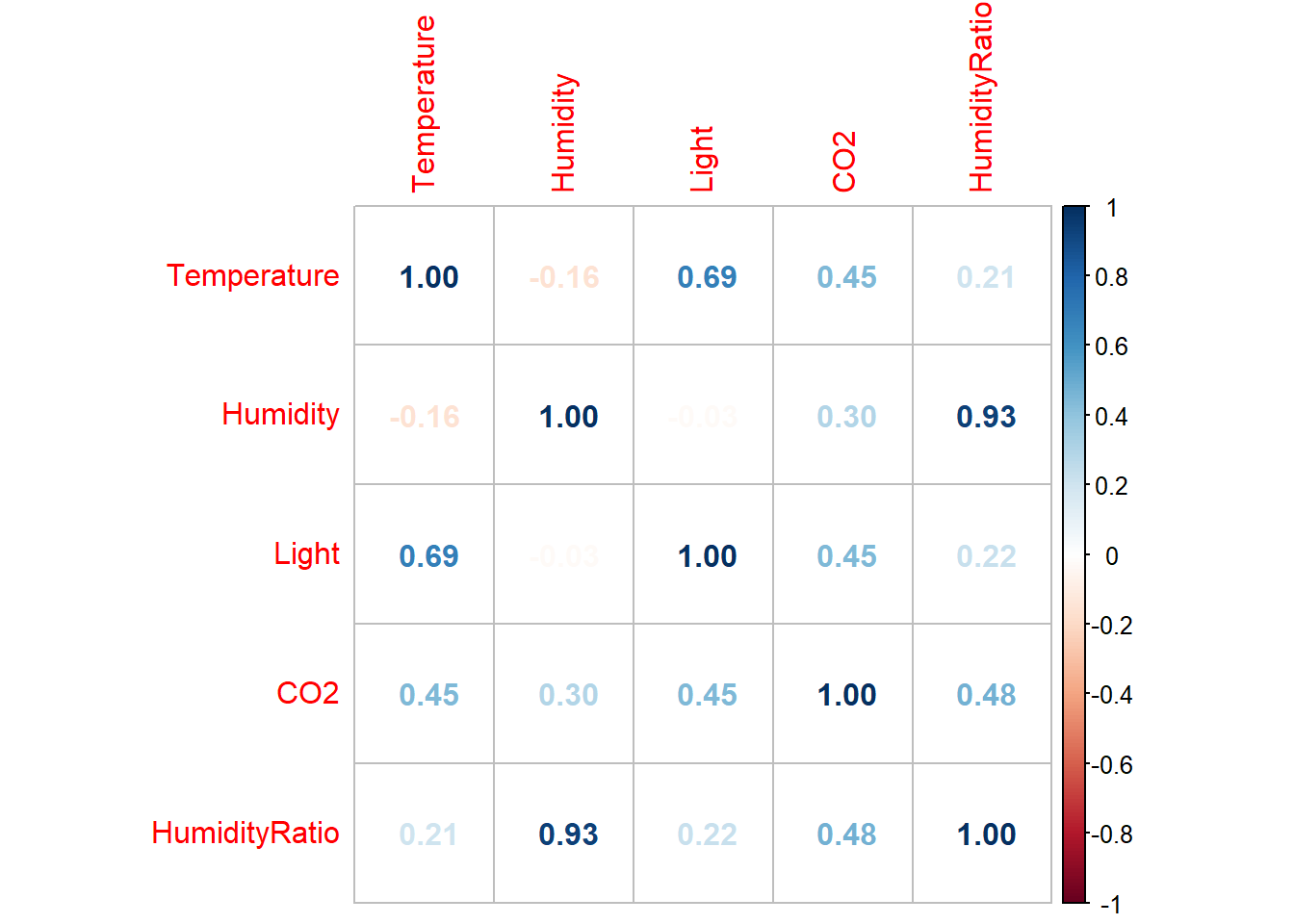
The support vector machine method (SVM) is a classification approach that performs well in a variety of settings and is intended for binary classification or regression. Primarily, its objective is to obtain a support vector to classify the data into groups and calculate the optimal decision boundary (Yu, et al). SVM learns from each input and predicts which of the classes the next data point belongs to (Shuaibu, et al) This approach is a generalization of a maximal margin classifier approach, which aims to find the hyperplane that has the farthest minimal perpendicular distance to the training observation. Being robust to individual observations and outliers, it is often considered one of the best “out of the box” methods for the classification of training observations (ISLR).

Extreme Gradient Boosting creates new learners that predict residual errors of its parent model sequentially to improve the overall performance of the model. Weak learners that have good performance have higher vote or weight in deciding the final ensemble model. The model optimizes loss functions by using the gradient descent which tries to find the direction of steepest, thus most efficient, descent that minimizes the error term of the model (Chong, Zak p131). Unlike a neural network, the weights are updated on the next predictor thereby learning from its mistake. is weight and being our learning rate and loss being .

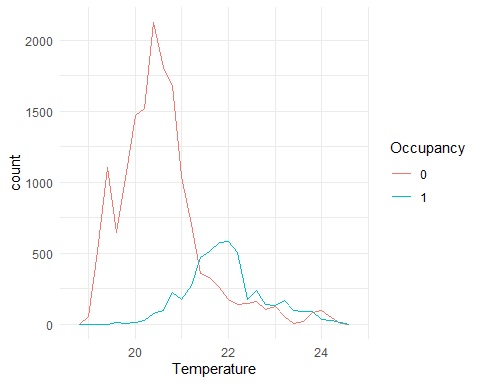
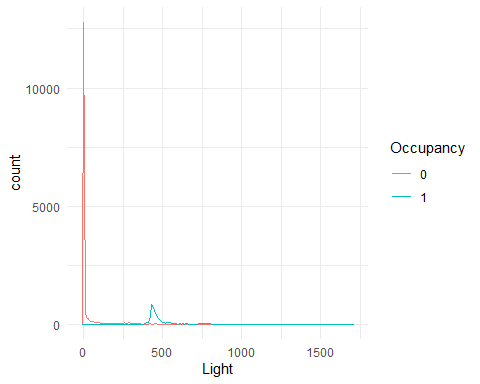
The xgBoost method is susceptible to overfitting so regulatory factors are included to improve performance, such as parallel data processing (Yu et al).

### Data Description

The Occupancy Detection dataset is obtained from the UCI Machine Learning Repository. Included in this dataset are 20,560 observations and 7 variables without missing variables. For this experiment, the timestamp variable, taken minute by minute, was removed so as to focus solely on each classification method’s predictive capabilities. Notably, there is a high association between Humidity and HumidityRatio, as seen in the figure below.



Since the data has been trained and tested using SVM and xgBoost, however, it is robust to multicollinearity and outliers. Moreover, the model conclusions indicated a stable output irrespective of the association.



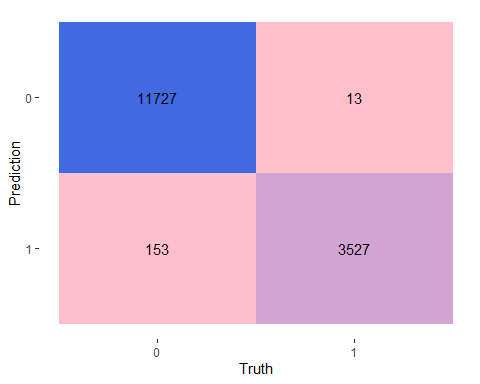
Based on the initial density plots, it appears that Light has a clear separation between occupancy states [0,1]. The initial inference is that Light will be a strong predictor. CO2 and Temperature seem to have some distinctions between occupancy states, while Humidity has no distinction.

#### **Methodology**

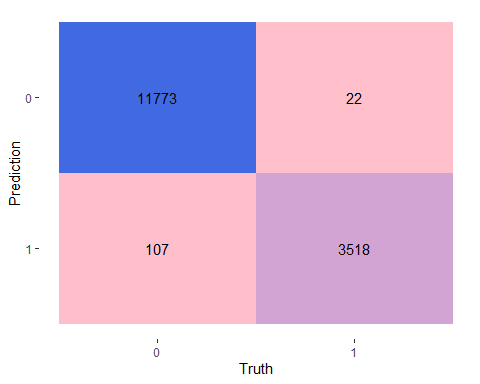
To compare the classification algorithms, the Occupancy Detection dataset was independently modeled by each method (support vector, boosted tree-based, and logistic regression as a general baseline), with an initial emphasis on accuracy of each classification model. The baseline approach of a binary logistic regression did not pass the Box-Tidwell test, thus not satisfying the linearity of log-odds assumption. For reproducibility, seed values were set. Confusion matrices, heatmaps, graphs, and other appropriate visualizations were used for a better understanding of the results. The data was split into training and testing sets. Ten fold stratified cross validation method was used for estimation. Laying hypercube sampling was used to tune the hyperparameters. Based on the results from the initial model, the Light variable was removed to test the predictive performance of the remaining variables using each classification method without a predominantly strong predictor. For more information, refer to the [GitHub](https://github.com/gnmergen/occupancy_detection) link (also found in the Appendix section).

#### **Results**

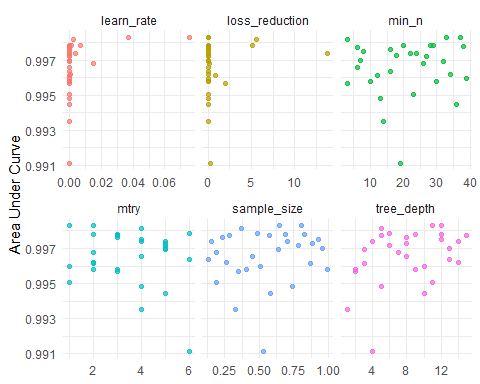
The SVM model correctly predicted 98.9% of variables. However, it was the lowest performing of the models used. The heatmap below shows the distribution of predictions on the training set.



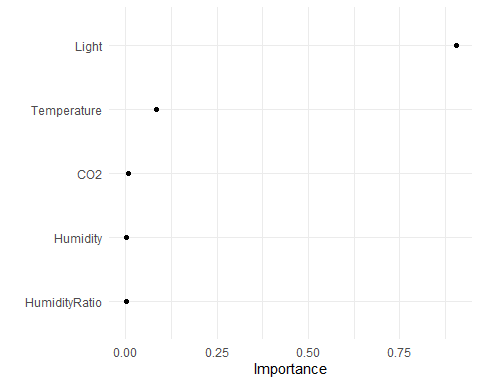
With xgBoost we obtained similar results on the training set



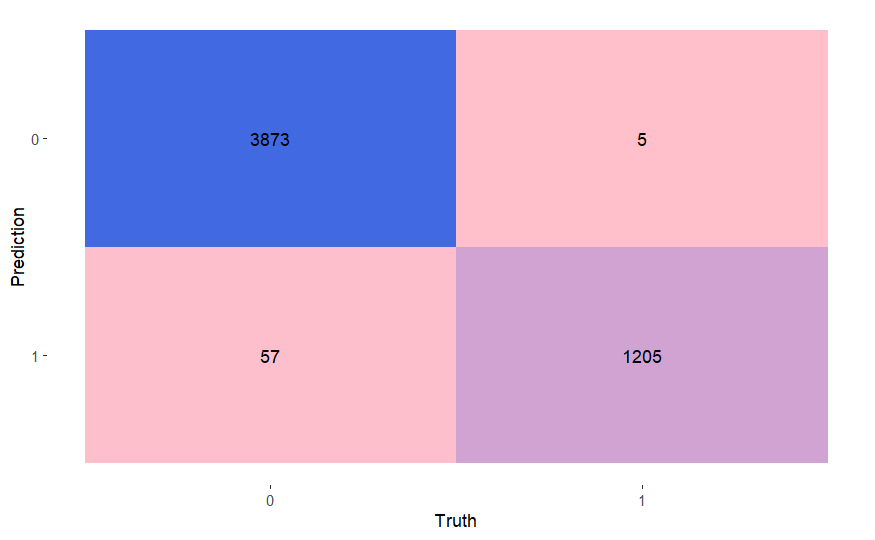
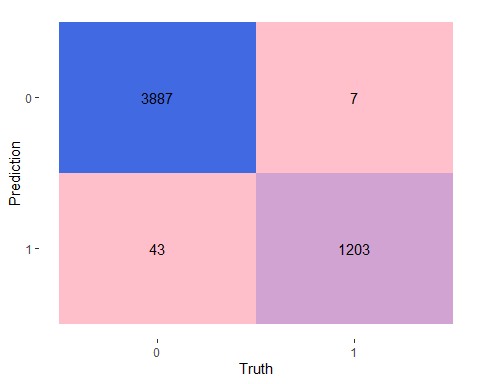
## Boosted Tree Model Specification (classification)  
## Main Arguments:  
## mtry = 4  
## trees = 1000  
## min\_n = 8  
## tree\_depth = 12  
## learn\_rate = 5.12308789188052e-09  
## loss\_reduction = 1.89494202573796e-10  
## sample\_size = 0.927796603294555  
## Computational engine: xgboost



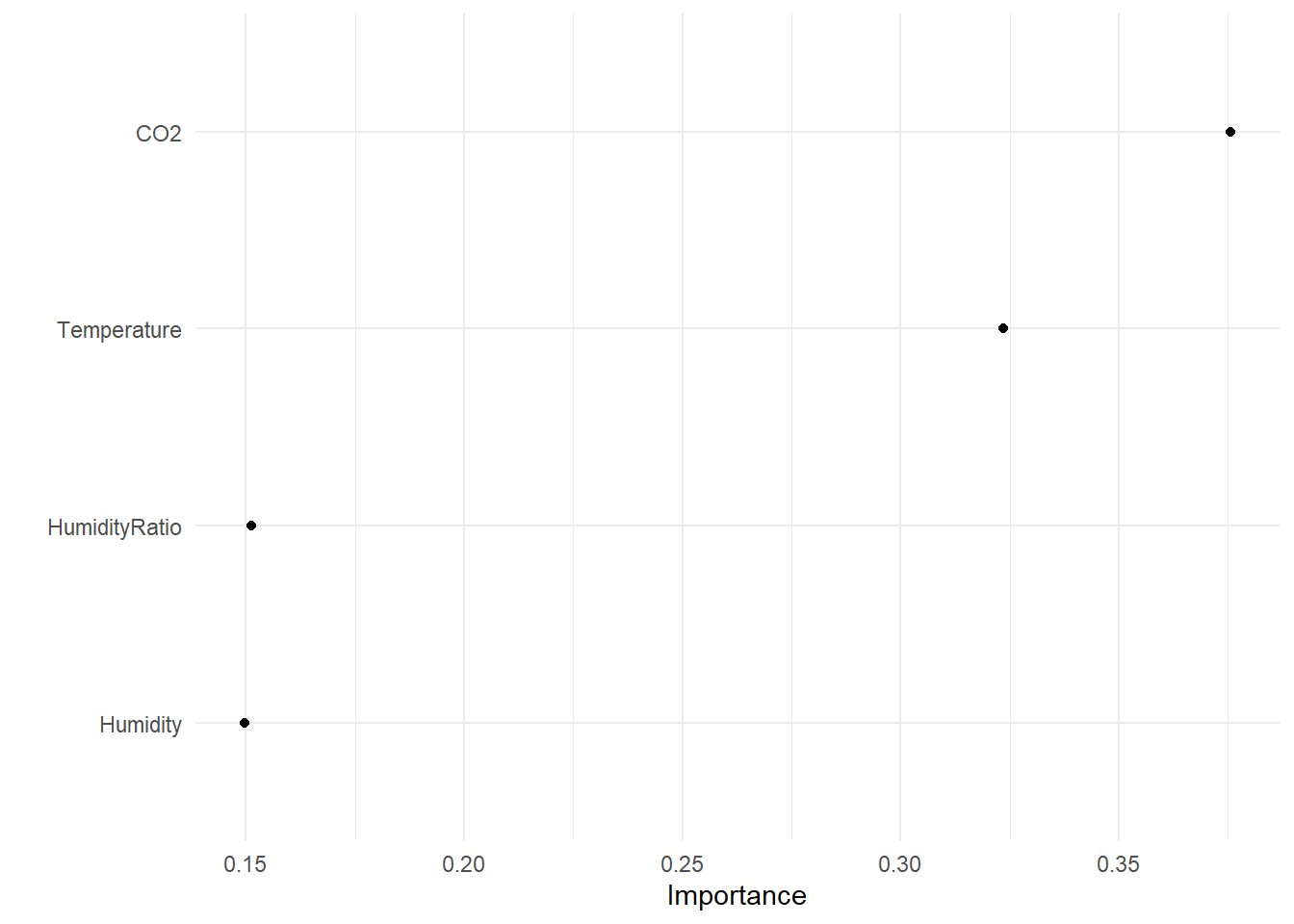
The model performance improved slightly using xgBoost compared to the SVM model with 99% accuracy . As previously mentioned, Light had the most predictive power out of all the variables.



Based on cross validation results, best xgBoost model was fitted to the testing data, resulting in 99.1% prediction accuracy, slightly better than SVM:

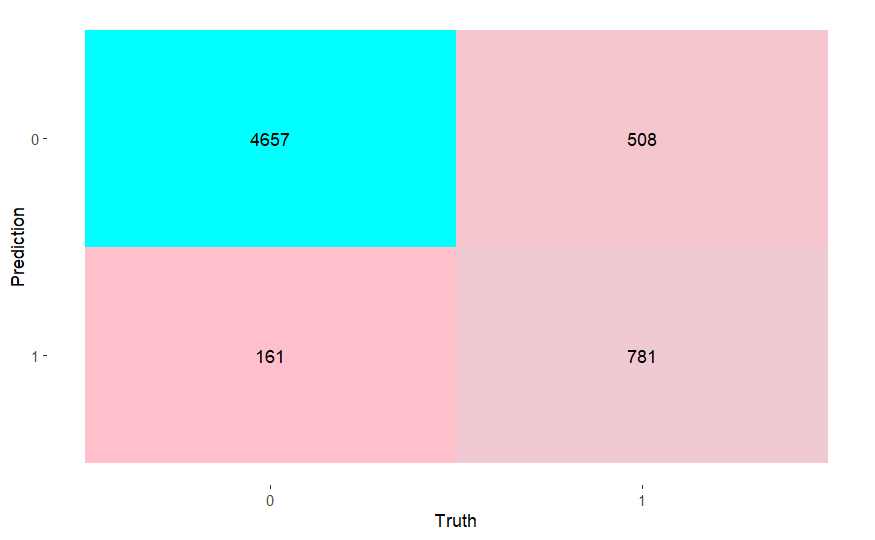
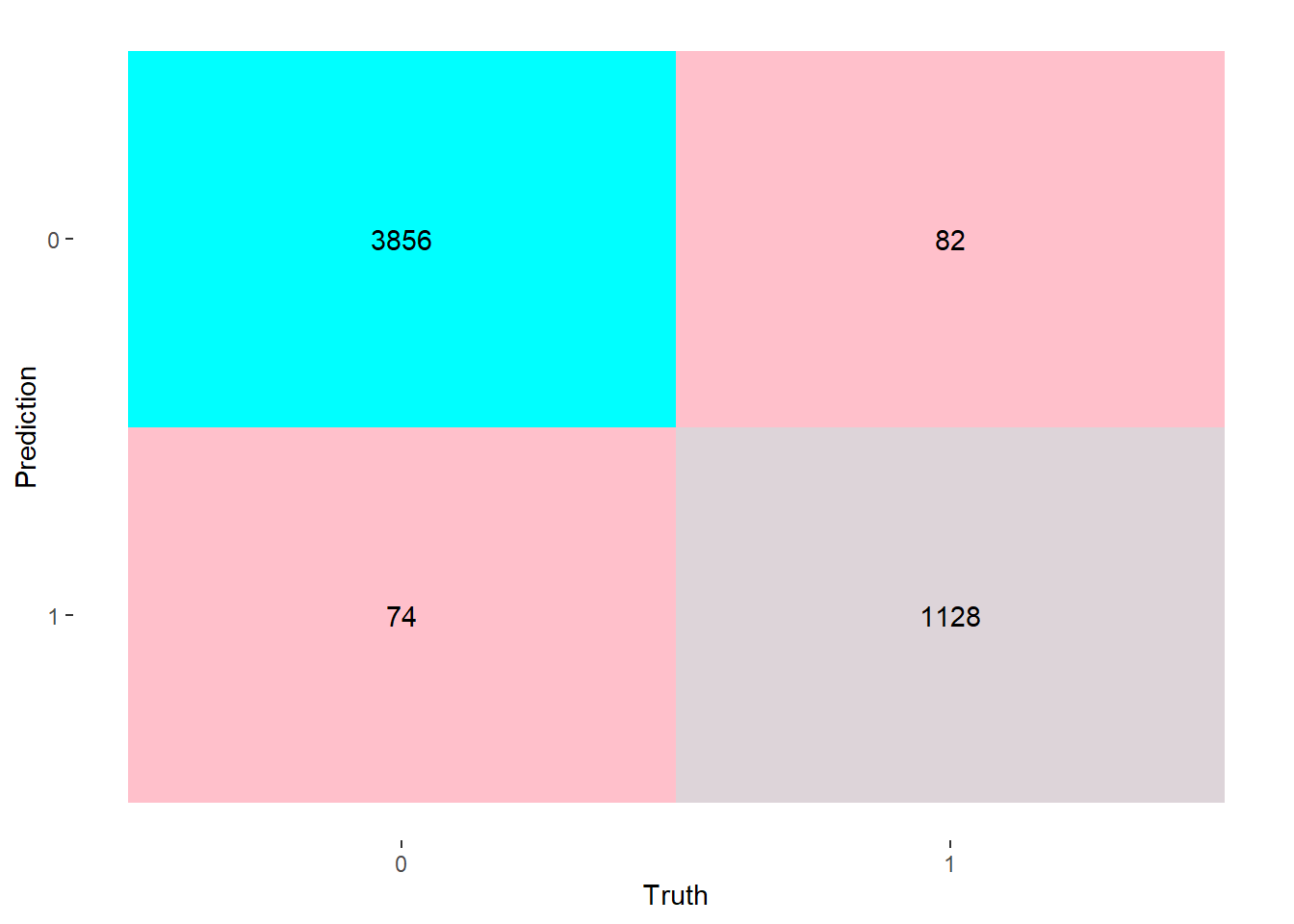
xgBoost Support Vector Machine  


Testing the predictive power of variables other than Light proved to be useful as we were able to achieve similar predictive accuracy. As seen from the below graph we are achieving close to 99% accuracy on the training dataset with CO2 being an important predictor.



We’ve also achieved 97% accuracy with xgBoost vs 89% with SVM on the testing set suggesting an absence of overfitting. The comparison of heatmaps is below.

xgBoost Support Vector Machine



### Conclusion

The svm and xgBoost models performed well, and similarly, on the Occupancy Detection dataset with an estimated error rate under 2% while using all predictors; Modeling without Light - xgBoost outperforms SVM. As a result of this study, it is determined that Light is not a sole indicator and strong functional predictions could be made if any one sensor fails. Important considerations include the calibration and sensitivity of each sensor and if that plays a role. Testing the models on a larger, scalable data set would be beneficial in attempting to further optimize the model, although the results would likely hold.

There are several possibilities for next steps and future research. Further research could explore the minimum and/or maximum values that trigger detection and determine the threshold of what classifies as a person, as opposed to a child or large animal. Alternatively, other machine learning methods, such as deep learning, could be applied.

**Data Source**

Luis Candanedo. (2016) “Occupancy Detection”. UC Irvine Machine Learning Repository. <https://archive-beta.ics.uci.edu/ml/datasets/occupancy+detection>

### References

Chen T., Guestrin, C. (2016). xgBoost: A Scalable Tree Boosting System. DOI: <https://dl.acm.org/doi/10.1145/2939672.2939785>

Edwin K. P. Chong, Stanislaw H. Zak (2013). An Introduction to Optimization (4th Ed.). John Wiley & Sons <https://books.google.com/books/about/An_Introduction_to_Optimization.html?id=8J_ev5ihKEoC>

Jakkula, V. (2006). Tutorial on support vector machine (svm). *School of EECS, Washington State University*, *37*. <https://course.ccs.neu.edu/cs5100f11/resources/jakkula.pdf>

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (*2nd ed.*)<https://hastie.su.domains/ISLR2/ISLRv2_website.pdf>

Shuaibu, A., et al (2020). Comparative Analysis of Diabetes Mellitus Classification Using Support Vector Machine and Extreme Gradient Boosting Algorithms. *International Conference on Electrical Engineering Applications (ICEEA'2020)*.<https://www.researchgate.net/profile/Nuraddeen-Ado-Muhammad/publication/348741664_Comparative_Analysis_of_Diabetes_Mellitus_Classification_Using_Support_Vector_Machine_and_Extreme_Gradient_Boosting_Algorithms/links/6012c1fca6fdcc071b9b2995/Comparative-Analysis-of-Diabetes-Mellitus-Classification-Using-Support-Vector-Machine-and-Extreme-Gradient-Boosting-Algorithms.pdf>

Yu, J., et al (2021). Forest Vertical Structure Mapping Using Two-Seasonal Optic Images and LiDAR DSM Acquired from UAV Platform through Random Forest, xgBoost, and Support Vector Machine Approaches. University of Seoul. DOI<https://doi.org/10.3390/rs13214282>

### Appendix

[Github Link](https://github.com/gnmergen/occupancy_detection)