Detecting Room Occupancy with Logistic Regression and GBM

Authors: Scott Amaranto, Zachary Gordon, Gaurav Rao

For this project, we sought to train a model aimed towards addressing the goal of improving energy efficiency in indoor settings. Machine learning can allow us to do this and help fight climate change. Internet Of Things (IOT) offers the ability to collect and analyze previously unseen data to achieve our goal. Built environments account for 40 percent of all carbon emissions. A chunk of those emissions are from utilities such as lights and HVAC. Turning off these utilities when a building is not in use can reduce emissions and save costs. Thus it is important to know when a room is in use. Building a machine learning model which can predict the occupancy of a room would allow appliances like lighting and HVAC to turn on and off automatically.

To train this model our group found a data set on Kaggle.com which contained reports from Internet Of Things (IOT) sensors, along with the ground truth of the room's occupancy. The dataset identifies whether or not there are 0, 1, 2, or 3 people in a room. 16 of the 18 features in the data are information from seven sensors (S1-S7). S1 through S4 contain information on room temperature in degrees Celsius, light levels in Lux, and sound levels in volts. S5 measured the Co2 levels of the room in parts per million and the slope of the change of Co2 levels. S6 and S7, identified in the data as PIR, were binary motion detectors which reported 1 if motion was detected and 0 if not. The sensors sent reports every 30 seconds over seven non-consecutive days in December of 2017 and January of 2018. The dataset contains the timestamps and dates of the reports. This comprehensive variety of data elements collected to measure different aspects of human activity was one of the main reasons we chose to move forward with this data set. Consideration and comparison of several categories of factors would allow for increased accuracy in predicting true room occupancy.

It was necessary to clean the dataset to achieve our goals. Our goal is to predict occupancy, not the number of people in the room. We mapped all samples with a non-zero value for "Room_Occupancy_Count" to 1 and all other samples to 0 in a new column we called "Occupied". We decided to not consider the date and time in any of our models. This is largely because it would turn the problem into a time-series problem which is outside the scope of our inquiry.

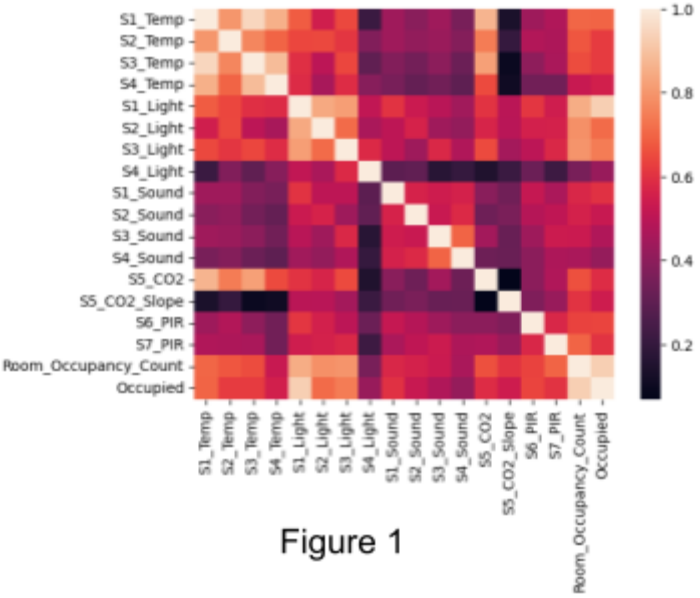

Figure 1

In the exploratory stage of our analysis we created distribution plots and looked at summary statistics of all of the features. We then looked for multicollinearity using a correlation heatmap as shown in **Figure 1**. The multicollinearity largely followed our expectations. The S1-S4 temperature, sound, and light features were moderately correlated with each other respectively. S4 temperature had the highest correlation with occupancy.

We decided to compare two models to find the best way to predict room occupancy. First we ran logistic regression, because it is considered a baseline for classification problems and has a lot of potential for inference. With logistic regression, we can compare beta values to determine how different features affect the data. The second model we chose was gradient boosted trees (GBM). GBM is regarded as a powerful model so we wanted to see if it could outperform logistic regression.

For both models we used an 80-20 train-test split and 3-fold cross validation for tuning hyper parameters. Using this split ratio allowed for a large portion of the data to be utilized for training while still maintaining a relatively sizable set of information to test the validity of our model without compromising on training. We chose 3 folds to balance runtime speed and having more validation sets. Having the same split and validation for both models ensures a fair and accurate comparison. To select parameters we used grid search to efficiently find the best parameters. Grid search is a common method for searching the parameter space. The set of parameters that produced the highest average accuracy between the different cross validations was chosen as the parameters for the model.

The only parameter we tested for logistic regression was the penalty. The grid search for this parameter was not strictly necessary, given we were only testing one parameter. However, we used it because it was needed for GBM. The penalty parameter determines how to penalize large coefficients. The larger the penalty the more small coefficients are favored. The penalty

creates a more generalized model and lowers the risk of over-fitting. We tested values of 0.001, 0.01, 0.1, 1, 2, 5, 10, 50, and 100.

We tuned three parameters for GBM: max depth, number of estimators, and learning rate. The max depth refers to how tall each of the trees can get, the number of estimators refers to the number of trees created, and the learning rate determines how quickly the predictor changes after each tree is generated. We tested 0.1, 0.2, 0.3, 0.5, and 0.7 for learning rate, 10, 50, 100, 300, and 500 for number of estimators, and 2, 3, 4, 5, 6 for max depth.

Logistic Regression

Testing Accuracy: 0.9417571569595261

	precision	recall	f1-score	support
0	0.93	1.00	0.97	1638
1	1.00	0.70	0.82	388
accuracy			0.94	2026
macro avg	0.97	0.85	0.89	2026
weighted avg	0.95	0.94	0.94	2026

Figure 2

GBM

Testing Accuracy: 0.9990128331688055

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1638
1	1.00	0.99	1.00	388
accuracy			1.00	2026
macro avg	1.00	1.00	1.00	2026
weighted avg	1.00	1.00	1.00	2026

Figure 3

Both models proved to be highly accurate. However, GBM had a higher accuracy with .999 compared to logistic regression's accuracy of .942. The naive guess of always predicting that a room is empty has an accuracy .812. It is clear that both models beat the naive guess. The precision, recall, and f1 scores for both logistic regression and GBM can be seen in **figures 2 and 3** respectively. GBM performs as well or better than logistic regression in all metrics. The grid search found that the optimal penalty for the logistic regression model was 50, which led to a mean validation accuracy of .999. **Figure 4** shows how the model accuracy changes with different penalty values. The optimal max depth for GBM was 3, the optimal number of estimators was 300 and the optimal learning rate was 0.1. The highest validation accuracy was .999. The heatmap in **figure 5** shows how different numbers of estimators and different learning rates affect the accuracy given a max depth of 3.

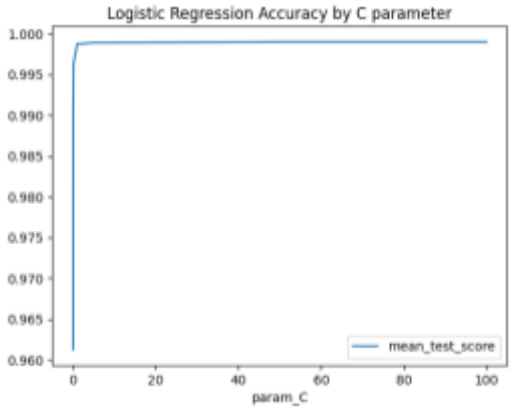

Figure 4

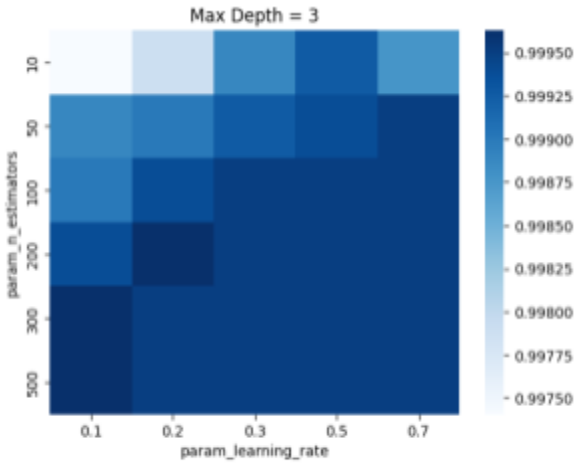

Figure 5

While overall GBM performs better than logistic regression, the regression still has value. It is easier to

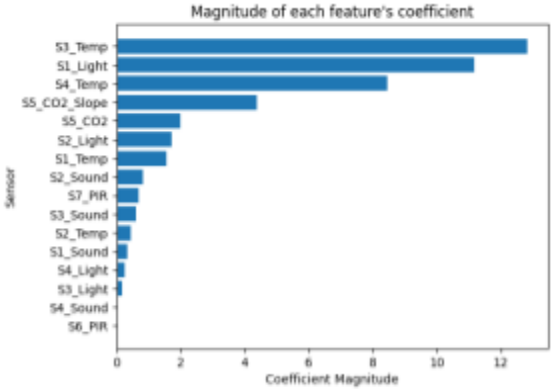

Figure 6

conduct inference with logistic regression, meaning we can more easily understand which features have the largest impact on the occupancy of the room. Because we standardized our features before running the logistic regression, we can compare the magnitude of the coefficients to see which features have larger effects. The larger the magnitude the larger the effect. The results can be seen in **figure 6**. The feature with the largest impact was S3 temperature and the feature with the lowest impact was S6 motion detection. We were surprised to see motion detection with such a low effect. However, temperature having a large effect makes sense. More information about the size and layout of the room is needed to fully explain this.

Our approach had a few identifiable limitations, such as the scope of considered sensor data. The purpose of our research, detecting occupancy in order to adjust light and temperature, were dependent on light and temperature sensors. We can see this in **figure 6**, where temperature and light from sensors were the 3 most important features for accurately predicting occupancy. We predict that time series analysis would be an important next step in establishing accurate temperature and light independent features. In addition, additional CO2 sensors could be beneficial to future datasets due to temperature/light independence and their outranking of most light and temperature sensors.

We were motivated towards this cause by a desire to leverage our knowledge and understanding of machine learning in a practical manner that still aligned with our concern for the increasingly negative effects of climate change. By developing a tool to improve energy efficiency, we aimed to make a small contribution to this overarching mission. A future endeavor could include developing a hardware solution to regulate light and heat in indoor spaces using our model for ensuring accuracy.

Bibliography

Github Repository -
https://github.com/zgordo/final-project

Room Occupancy Estimation Data Set, by Ananth R. -
https://www.kaggle.com/datasets/ananthr1/room-occupancy-estimation-data-set