**Anyone There?**

Joshua Dow, [jdow@bellarmine.edu](mailto:jdow@bellarmine.edu)

Drew Buhr, [dbuhr@bellarmine.edu](mailto:dbuhr@bellarmine.edu)

**ABSTRACT**

Can one predict whether someone is occupying a room or not using only scientific measurements? We can. Given a data set that measures multiple environmental variables of a room and if it is occupied or empty, we are able to predict the occupancy of the room to extreme precision given only the measurements. Through the entire data analysis process of cleaning and manipulating our data to a useful form, we use machine learning to generate our predictions. Creating and running different machine learning models, we are able to build a model the predicts to a 99% accuracy. Given certain measurements we can tell if anyone is there.

1. **INTRODUCTION**

They data set we were assigned is the collection of a room’s temperature, humidity, light, CO2 levels, humidity ratio, and occupancy of the room every minute for 17 days in February. Temperature is measured in Celsius. The humidity level is given as a percentage. In lux is how light is collected in the data set. The CO2 level of the rooms is measured in ppm, part per million. The humidity ratio was calculated using a derived quantity from temperature and humidity in kilograms of water-vapor over kilograms of air. Given these variables of the room, our goal is to predict whether the room is occupied or not.

1. **BACKGROUND**

According to the summary on https://archive-beta.ics.uci.edu/ml/datasets/occupancy+detection, the data set was created for practice. With the occupancy data set, the main purpose of the data set was to collect data every minute measuring a room’s environment such as temperature, light levels, CO2 levels, and humidity and see if environmental measurements can determine the occupancy of a room.

1. **EXPLORATORY ANALYSIS**

This data set is 20,560 samples with 7 variables of different data types. Through evaluation we found this data set to not be missing any data entries and there was no significant outliers. The pairplot we ran confirmed our expectations of the relationship between the variables, like Humidity and HumidityRatio would probably be heavily correlated. We did find however the variables Light, CO2, and Occupancy to be skewed, but that was to be expected since the measurements would be drastically different when no one is in the room compared to when someone is.

We found this heatmap and the density plots of Temperature, Light, and CO2 to be very helpful in pointing out correlations among our variables. To no surprise, light had the greatest correlation to Occupancy since most people turn on the lights when they enter a room. We also expected that Temperature and CO2 would be correlated to Occupancy since people give off heat and breath out CO2, so we expected the measurements to be higher than when the room is unoccupied. Likewise, Humidity and HumidityRatio are highly correlated. Shown below is the heatmap and the temperature, light, and CO2 density plots.

Square

Description automatically generated

**Chart, line chart

Description automatically generated**

**Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated**

**Table 1: Data Types**

|  |  |
| --- | --- |
| *Variable Name* | *Data Type* |
| V1 Date | Object |
| V2 Temperature | Float |
| V3 Humidity | Float |
| V4 Light | Float |
| V5 CO2 | Float |
| V6 HumidityRatio | Float |
| V7 Occupancy | Integer/ switched to Object |

1. **METHODS**
   1. *Data Preparation*

When it came to preparing the data, we found very little cleaning needed to be done. We basically found no missing data or significant outliers, so we did not have to do a lot of data manipulation. Our given data units was really easy to work with, so we decided to keep the units the same throughout our examination. Only when we began our prediction did we decide to normalize the data set and see if that impacted the accuracy of the model. However, the date and time of the measurements was not found to be very useful, therefore we decided to drop that column. We also decided to manipulate the ‘Occupancy’ column to “yes” or “no” instead of “1” and “0”, respectively, because we believed it would be easier for us to remember and it did not impact the prediction model. The biggest speedbump we found ourselves running into was creating a potential bias by merging the different given data sets into on big data set and then running our models instead of using the given testing and training data sets.

* 1. *Experimental Design*

Table 2: Logistic Regression Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | All of the variables without ‘Date’ using the first given splits for training and testing. Roughly 10% of the data used for testing |
| 2 | All of the variables without ‘Date’ using the second given splits for training and testing. Roughly 45% of the data used for testing |
| 3 | All of the variables without ‘Date’ normalized using the first given splits for training and testing. Roughly 10% of the data used for testing |
| 4 | All of the variables without ‘Date’ normalized using the second given splits for training and testing. Roughly 45% of the data used for testing |
| 5 | The square root of all of the variables without ‘Date’ using the first given splits for training and testing. Roughly 10% of the data used for testing |
| 6 | The square root of all of the variables without ‘Date’ using the second given splits for training and testing. Roughly 45% of the data used for testing |

We found our data to be very predictable, as you’ll see in the summaries below. All of our Logistic Regression experiments returned an accuracy of 98% or 99%.

A picture containing text, receipt

Description automatically generated

A picture containing text, receipt

Description automatically generated Experiment I Experiment II

A picture containing text, receipt

Description automatically generated

A picture containing text, receipt

Description automatically generated Experiment III Experiment IV

Table

Description automatically generatedTable

Description automatically generated

Experiment V Experiment VI

Table 3: KNN Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | All of the variables without ‘Date’ using the first given splits for training and testing. Roughly 10% of the data used for testing |
| 2 | All of the variables without ‘Date’ using the second given splits for training and testing. Roughly 45% of the data used for testing |
| 3 | All of the variables without ‘Date’ normalized using the first given splits for training and testing. Roughly 10% of the data used for testing |
| 4 | All of the variables without ‘Date’ normalized using the second given splits for training and testing. Roughly 45% of the data used for testing |
| 5 | The square root of all of the variables without ‘Date’ using the first given splits for training and testing. Roughly 10% of the data used for testing |
| 6 | The square root of all of the variables without ‘Date’ using the second given splits for training and testing. Roughly 45% of the data used for testing |

We found our KNN model to be successful in the prediction of the occupancy of the room because all of the experiments, from Table 3, returned an accuracy score of 96% or higher. The classification reports are shown below.

*Table

Description automatically generatedTable

Description automatically generated* Experiment I Experiment II

Table

Description automatically generatedA picture containing text, receipt

Description automatically generated

Experiment III Experiment IV

A screenshot of a computer

Description automatically generated with low confidenceA picture containing text, receipt

Description automatically generated

Experiment V Experiment VI

1. *Tools Used*

We used Jupyter Notebook for Lenovo Chromebook to analyze this data set. In addition to Python v3.5.2 we used Pandas 0.18.1, Numpy 1.11.3, Matplotlib 1.5.3, Seaborn 0.7.1, and SKLearn 0.18.1. We chose to use these tools because we found Python to be easier for us to use and understand, especially compared to R. We used the additional tools to help with calculations and plotting of the data. For example, we are comfortable and like the layout of Seaborn plots.

1. **RESULTS**
   1. *Classification Measures*

The classification of how successfully our models were is shown in the reports above. We found both of our models to be extremely successful since they predicted the correct occupancy with extreme precision.

* 1. *Discussion of Results*

If you have to force us, we will conclude that our Logistic Regression model is better than our KNN model. Since our data set is very cut and dry, pretty linear, it is no surprise that the Logistic Regression model is so successful. Since there is some overlap in some of the measurements of when a room is occupied or not, most likely occurring most when someone just entered or left the room near the minute mark, the KNN model will sometimes predict inaccurately since the KNN model groups clustered data to labels. Thus, the model would struggle when measurements would be very close in value but occupancy would be different.

* 1. *Comparison of Models*

Our Logistic Regression model was better than our KNN model. Since a Logistic Regression model predicts our format of data better than a KNN model, explained in further detail above in “Discussion of Results”, it is not surprising the Logistic Regression model predicts at a higher accuracy.

* 1. *Problems Encountered*

The biggest problem we had with this data set was when we first ran our models, we merged the given training and testing data into one big data set and then ran our experiments. We found essentially perfect accuracy with all of our experiments. Since that is almost impossible, we took a step back and reread our code line by line and found that merging the data sets given by the author was probably what was causing the problem. We were correct quickly adjusted our models and experiments.

* 1. *Limitations of Implementation*

We believe Logistic Regression is the best model to run our data set since our data set is pretty linear. Proof that it is the best model for our data is shown the high accuracy of 98% or better in all of our experiments. Our KNN model still predicted the occupancy of the room very well, but not as good as the Logistic Regression model. As stated before, the KNN model potentially struggled with the overlapping of similar measurements even though the occupancy of the room was different. This problem most likely occurring in the minutes when someone just entered or left the room.

* 1. *Improvements/Future Work*

For the future, we would consider possibly removing variables such as HumidityRatio and see if that significantly impacts our model. We were curious during the evaluation whether we needed both Humidity and HumidityRatio, but we decided to keep both variables since our data set was not too big.

1. **CONCLUSION**

With the right environmental variables, temperature, light, CO2 levels, humidity levels, and the humidity ratio, we are able to very accurately predict whether someone is there or not. Our Logistic Regression model predicts almost flawlessly no matter how or how much training the model receives. It was very successful given the raw, normalized, and square root measurements no matter if it trained on nearly 90% or 50% of the data set. Even though outshined, our KNN model was very successful in its predictions as well. Predicting at an accuracy of 96% or better every experiment. Overall, both of our models worked tremendously well in accurately predicting the occupancy of the room.

**REFERENCES**

List any websites, books, articles, etc. that you found useful while you worked on this project. It is not necessary to cite the references in the paper unless you specifically mention it in the text.POST ANY YOU USED!!!

*Division of Labor*

Due to Drew’s wild baseball schedule, Joshua did an amazing job on essentially all of the coding in Python. We collaborated briefly on the code at the beginning of setting up our models, but after we figured out the right training and testing data sets Joshua finished the coding. Drew contributed to essentially all of the paper analysis. We were able to meet three times and go over the code and analysis together, so neither one of us was in the dark to any of the project.

**DELETE LATER!!! Other directions:**

1. 10-pt, Times New Roman, 1” margins all around (if you use this template you are already set).
2. **Ensure all tables and figures are numbered appropriately and referenced in the text.** See examples above and below.

|  |  |
| --- | --- |
| **Figure 1: Comparison of X/Y from dataset (single plot) (8 pt.)** | **Figure 2: (a) Function Output (b) A against B (multiple plots) (8 pt.)** |