Project 4

Anton Stråhle & Max Sjödin

28 december 2020

Introduction

This is a project for the course MT7038 which was given during the fall of 2020.

Data

We picked the `Occupancy Detection` data set as we wanted to work with a binary classification problem as this would allow us to apply most of the methodologies discussed throughout the course. As there are quite a few different binary data sets at UCI we specifically chose our data set as it had a sizeable number of instances as well as few, but intuitivley explanatory, features.

Attributes

The data `Occupancy Detection` data set includes snapshots of a specific room every minute throughout the course of about of several days. The aim is to classify the current `Occupancy` of the room using the five features, `Temperature`, `CO2`, `Humidity`, `HumidityRatio` and `Light`, which are observed each minute. The first three features are quite self-explanatory but the two final ones could use some clarification. The `Light` in the room is the light intensity, measured in Lux whilst the `HumidityRatio` is vaguely described as a "derived quantity from temperature and relative humidity, in kgwater-vapor/kg-air".

Exploration

From a quick overview we see that the data set is quite unbalanced.

Table 1: Unoccupied

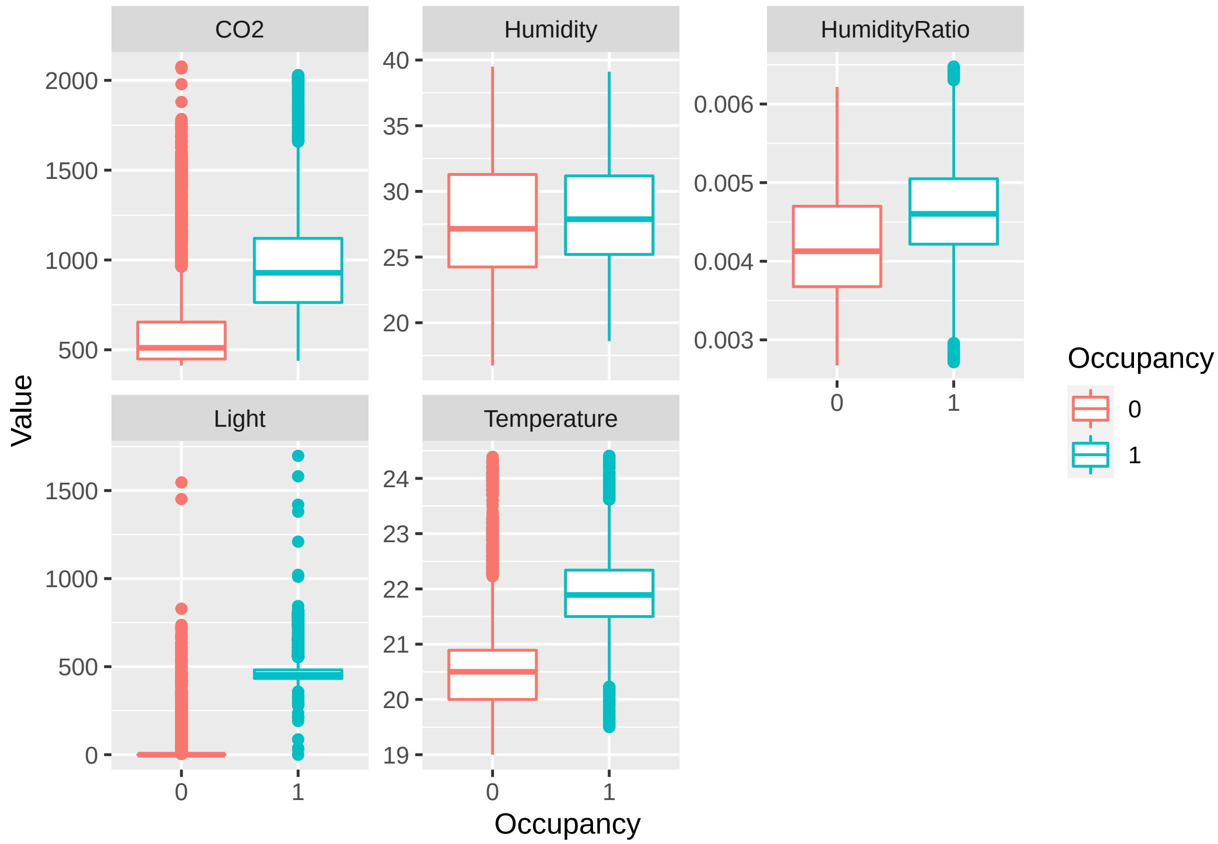
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
Temperature	1	15810	20.585	0.895	20.500	20.488	0.741	19.000	24.390	5.390	1.325
Humidity	2	15810	27.530	5.119	27.150	27.556	5.411	16.745	39.500	22.755	-0.030
Light	3	15810	25.238	81.824	0.000	2.966	0.000	0.000	1546.333	1546.333	4.667
CO2	4	15810	604.997	253.027	511.000	545.863	102.299	412.750	2076.500	1663.750	2.442
HumidityRatio	5	15810	0.004	0.001	0.004	0.004	0.001	0.003	0.006	0.004	-0.151
Occupancy*	6	15810	1.000	0.000	1.000	1.000	0.000	1.000	1.000	0.000	NaN

Table 2: Occupied

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
Temperature	1	4750	21.976	0.818	21.890	21.938	0.619	19.500	24.408	4.908	0.414
Humidity	2	4750	28.076	4.472	27.882	28.040	4.305	18.600	39.117	20.517	0.156
Light	3	4750	481.967	94.704	454.000	461.597	31.135	0.000	1697.250	1697.250	2.953
CO2	4	4750	975.322	317.261	928.583	943.895	267.486	439.000	2028.500	1589.500	0.997
HumidityRatio	5	4750	0.005	0.001	0.005	0.005	0.001	0.003	0.006	0.004	-0.059
Occupancy*	6	4750	2.000	0.000	2.000	2.000	0.000	2.000	2.000	0.000	NaN

We have about four times more unoccupied than occupied minutes. As the data is observed around the clock it is of course natural that the room is unoccupied during a majority of the day. Due to this quite severe inbalance we have to make sure that our training, validation and testing sets reflect this inherent property of the data. As such we decided to concatenate the three provided data sets (training, validation and testing) and split these up into balanced sets ourselves as the data providers seems to have just split the complete date by the timestamp which leads to a quite severe inbalance as a weekend is included (i.e. no Occupied observations for two days).

We also chose to standardize to allow for better performance of scale dependent classifiers, e.g. kNN or SVM. As some of the features include some very major deviations as can be seen in the figure below this choice to not standardize would surely impact the performance of these classifiers negativley.



As can be seen in both the figure and the tables above the feature `Light` seems to to quite well in describing the current occupancy of the room. This is quite evident as people rarely gather in a room with the lights turned off, and (hopefully) turn the lights of when they leave. This features solely dominates the others when it comes to classification and is ,at least according to us, quite boring. As such we'll choose to excluded it which will be discussed under **Analysis** in order to actually be able to perform a somewhat comprehensive analysis that doesn't just include the questions *Was the lights off?*.

Methodology

As we're dealing with a binary classification problem there are some major routes that we decided to test out. Those being SVM, KNN, Decision Trees and Logistic Regression. Our idea is to examine all methods quite shallowly and then go a bit deeper for one or two that shows promise.

TÄNKER ATT VI SKRIVER VARFÖR METODERNA ÄR RIMLIGA OCH DÄREFTER GÅR IGENOM KORT VAD VI FÅR FÖR BROAD RESULTS (ALLTSÅ INNAN VI FÖRFINAR MODELLERNA). SEN SKRIVER VI OM EN/TVÅ BÄSTA UNDRER ANALYSIS DÄR VI GÖR DJUPGÅENDE ANALYS TYP.

SVM

KNN

Decions Trees

Logistic Regression

Analysis

Bibliography

Appendix