Predicting shop floor occupancy – Student Guide

# Overview

Contoso Corporation is a construction company building shop floors for many manufacturers across the United States. They are interested in incorporating sustainability into their shop floor design as they embark on smart building development. To support this effort, Contoso has invested in IoT technology, installing sensors on 51 of their shop floors. As a launching ground for their sustainability journey, Contoso needs to understand the data collected by their sensors and the factors influenced by the occupancy rate. Using this knowledge, they want to regulate temperature, humidity, light, and monitor CO2 levels on their shop floors.

**This hackathon enables attendees** to access the data Contoso has collected from their sensors to predict if a room does or does not have any occupants. The dataset contains a week’s worth of records or each of the 51 shop floors. This data includes information about CO2 concentration, humidity, room temperature, and luminosity.

The dataset has been obfuscated to prevent any leak of IP or identities from prior analysis, and thus the column variables will be general in nature.

**During the “hacking” attendees will focus on** understanding the data, searching for trends, exploring correlations, the characteristics and impact between C02, temperature, humidity, light and occupancy rate. Hackers will determine which of these collected features could potentially help in determining if a shop floor is occupied.

**This hackathon enables attendees to** help identify which type of data collected could prove most beneficial in the development of smart buildings.

**This hackathon simulates a real-world scenario** where a construction company is striving to incorporate sustainability in future shop floor designs.

**By the end of the hackathon**, attendees will have built a classification model to predict if a shop floor is occupied.

The main goal for Contoso Corporation is to incorporate sustainability to design smart Carbon and Energy efficient buildings. The challenge is to analyze and interpret data collected and understand the factors influenced by occupancy rate for Contoso.

Once all hackathon challenges are completed, you should be able to:

* Create an Azure Machine Learning workspace
* Collect and process the measurements from the shop floor sensors for data exploration
* Identify the occupancy of the floor with the help of a predictive model
* Understand the features impacted by the shop floor occupancy and help prevent casualties

# Prerequisites

In these challenges, you will provision an Azure Machine Learning workspace and you will then use the Compute Instance to explore data interactively.

## Prerequisites for Tools

To complete the challenges, you will need to have background knowledge of the following:

* **Language:** Python
  + **Tools:** Azure Machine Learning Studio (Auto ML, Designer can also be used for this challenge but not necessary); Jupyter notebooks
* **Basic Data Science Knowledge:** Data Exploration and Classification modeling building (Relevant prep materials: [Create machine learning models - Learn | Microsoft Docs](https://docs.microsoft.com/en-us/learn/paths/create-machine-learn-models/))

Other basics would be:

* A web browser
* A Microsoft account
* A Microsoft Azure subscription – that would be provided
* A Windows, Linux, or Mac OS X computer
* The challenge files for this course

## Setting up the system

To complete the challenges, you will need the following:

A Microsoft Azure subscription. If you do not have a subscription, please connect with your proctor who will be able to provide one.

An Azure Machine Learning workspace. If you do not have an Azure Machine Learning workspace in your Azure subscription, follow these steps to create one:

1. Sign into the Azure portal using the Microsoft account associated with your Azure subscription.
2. Select ＋Create a resource, search for Machine Learning, and create a new Machine Learning resource with the following settings:
   1. **Workspace Name**: enter a unique name of your choice
   2. **Subscription**: your Azure subscription
   3. **Resource group**: create a new resource group with a unique name
   4. **Location**: choose any available location
3. Wait for your workspace resource to be created, this can take a few minutes. Go to your workspace in the portal, on the Overview page launch Azure Machine Learning studio (or navigate to <https://ml.azure.com>), and sign in using your Microsoft account.
4. In Azure Machine Learning studio, toggle the ☰ icon at the top left to view the various pages in the interface. You can use these pages to manage the resources in your workspace.
5. Create a compute instance, you will need a compute instance in your Azure Machine Learning workspace to run this exercise.

In Azure Machine Learning studio, view the Compute page for your workspace (under Manage). On the Compute Instances tab, if you already have a compute instance, start it; otherwise create a new compute instance with the following settings:

1. **Virtual Machine type**: CPU
2. **Virtual Machine size**: Standard\_DS11\_v2
3. **Compute name**: enter a unique name
4. Wait for the compute instance to start, this may take a few minutes

## Links & Resources for Post Learning Recommendations

* + 1. Sustainability Resources:
       - * Sustainability FAQs: [FAQ (sharepoint.com)](https://microsoft.sharepoint.com/sites/sustainability/sitepages/faq.aspx)
         * MSX Content: <https://aka.ms/MSUSSustainability>
* Sustainability Hub: <https://microsoft.sharepoint.com/sites/sustainability/>
* Environmental Priorities- Carbon: <https://microsoft.sharepoint.com/sites/sustainability/SitePages/Program-Carbon.aspx>
* Environmental Priorities-Ecosystems: <https://microsoft.sharepoint.com/sites/sustainability/SitePages/Ecosystems.aspx>
  + 1. Data Science:
       - * Azure Machine Learning: <https://docs.microsoft.com/en-us/azure/machine-learning/>
         * Data Exploration and Model building: [Create machine learning models - Learn | Microsoft Docs](https://docs.microsoft.com/en-us/learn/paths/create-machine-learn-models/)

# Challenges

There are 4 challenges in total in this for you to successfully complete this hackathon, which are listed below:

## Challenge 1: Load the data set to Azure Machine Learning Workspace

A sophisticated data-collection device, the sensor is a crucial and fascinating component of the Internet of Things (IoT). The purpose of sensors is to collect analog data from the physical world and translate it into digital data assets. Sensors are measuring just about any aspect of the physical world. The calibration of sensors allows them to be tailored to application-specific functions. In this dataset, sensors have been calibrated to measure temperature, humidity, CO2 concentration, luminosity and PIR (motion detection) with accuracy. This sensor data is tasked with capturing information relevant to a shop floor design, so the data can be used to make process improvements for the purpose of increasing carbon and energy efficiency in shop floors.

### Objectives

* How is sensor data collected in Azure Blobs?
* How to ingest and wrangle the data to generate insights from it?

### Tasks

* Data Ingestion Task 1: Download the data from Azure Blob Storage
* Data Ingestion Task 2: Read that data into a single dataframe

### Azure Blob Data Storage Details:

* Connection String:
  + "DefaultEndpointsProtocol=https;AccountName=sustainabilityhackathon;AccountKey=Ejq44H9MM9EZj45ly40vT1cHsZmUAjaIRR+KE5jyqRBqZ+QRZYXwB4+0lJNOGQlHQMVSACChVv9n2GovIPl/WA==;EndpointSuffix=core.windows.net"
* Blob Container:
  + csv

Hint 1: Some of the data are collected every 5 seconds, but the PIR motion sensor data are only collected every 10 seconds. If you merge or aggregate the data, make sure you account for data with null values.

Hint 2: The UNIX timestamp in this data measures the number of seconds since 1970-01-01 00:00:00. When converting to a datetime from UNIX, make sure the conversion formula has a unit of seconds and not nanoseconds.

Hint 3: When you collect the data from each shop floor, add a column for the shop floor number so that is recorded in the data.

**Data Description**There are 5 types of measurements from each shop floor sensor. Data is collected over a period of one week from Friday, August 23, 2013 to Saturday, August 31, 2013.

* co2.csv - Carbon-dioxide concentration (sampled every 5 seconds)
* humidity.csv - humidity (sampled every 5 seconds)
* light.csv - luminosity (sampled every 5 seconds)
* pir.csv - PIR (passive infrared) motion sensor data (sampled every 10 seconds)
* temperature.csv - shop floor temperature (sampled every 5 seconds)

Each dataset has UNIX Epoch – It is the number of seconds that have elapsed since January 1, 1970. You can choose to convert it to datetime.

**Acknowledgement**: It is hereby acknowledged that the data used here was sourced from publicly available files and channels.

## Challenge 2: Data Exploration

Data exploration is an approach to understand what is in a dataset and the characteristics of the data. These characteristics can include size or amount of data, completeness of the data, correctness of the data, possible relationships amongst data elements or files/tables in the data. Data Exploration is aimed at understanding the nuances of the data, and defining basic metadata (statistics, structure, relationships) for the data set that can be used in further analysis. Once this initial understanding of the data is had, the data can be pruned or refined by removing unusable parts of the data, correcting poorly formatted elements and defining relevant relationships across datasets.

### Objectives

* What are the ways data can be pruned or refined by removing unusable and poorly formatted data portions?
* How to visualize relevant relationships amongst features across dataset?

### Tasks

### Data Exploration Task 1: Is there any missing data? Does data need any imputation and take the suitable data imputation measures if needed.

### Data Exploration Task 2: Do certain days and hours of the day have more occupancy? Any interesting trends you can comment on?

### Data Exploration Task 3: Identify correlations between the features. Are there feature that are highly correlated to shop floor occupancy rate?

### Data Exploration Task 4: What are the characteristics of the features CO2, temperature, humidity and light? Hint: Take a look at the distribution of each feature. Are the values concentrated around a features mean or median or are several values vastly different than the mean or median? Are the mean and median the same, or is there a gap between them?

### Data Exploration Task 5: Which of the features CO2, temperature, humidity and light could potentially help in determining if a shop floor is occupied or not? Hint: The PIR values in the data have different measurements for different levels of occupancy. It would make sense to create a new variable denoting if PIR was 0 or greater than 0.

## Challenge 3: Build a classification model to predict if a shop floor has occupants or not

A machine learning model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data. Once you have trained the model, you can use it to reason over data that it hasn't seen before and make predictions about those data.

### Objectives

* What machine learning models are suitable for this and similar problem statements?
* How to interpret how the machine learning models are tuned to predict future unseen data?

### Tasks

### Classification model Task 1: What additional features might be interesting for this problem? Hint: Look back at Challenge 2, Task 2. Are there any trends regarding any features other than the ones you analyzed from Challenge 2, Task 5 (for example, hour and date)? Would those features be numeric or categorical?

### Classification model Task 2: Choose a model and a metric to build a model to predict occupancy. Hint: How well would the model perform on each class? How often would it correctly predict that the floor is occupied? How often would the floor occupied when the model predicts that it is? Is there a metric that can answer both of those questions?

### Classification model Task 3: Explain which features are most useful for predicting whether a shop floor is occupied. Explain the relationships between the different model features and model predictions. Hint: Some Python packages will have built-in methods for determining feature importance.

## Cleaning up

If you used a compute instance in an Azure Machine Learning workspace to complete the exercise, use these steps to clean up:

* Close all Jupyter notebooks and the Juptyer home page
* In Azure Machine Learning Studio, on the Compute page, select your compute instance and stop it

# Value Proposition

* Develop fluency in sustainability topics especially in carbon and energy space to have meaningful conversations with customers and partners
* Join the community of Microsoft Sustainability Champions and get badged

# Technical Scenarios

* Environmental Monitoring​
* Energy consumption / embodied carbon monitoring and reporting​
* Facilities/equipment operating efficiency optimization​
* Energy source optimization

# Audience

* + - Facilitator/Proctor: ATU, CSU, STU, SMC, OCP, CSA
    - Student: CSA, CSU

# Market Roadmap

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Goals** | **Opportunity Areas** |
| **Carbon /  Energy** |  | * Reduce total & per capita energy consumption * Reduce GHG emissions, Scope 1-3 * Increase contribution of Green Power in total power consumption * Energy & grid management | * Monitor power consumption across operations * Monitor CO2 emissions across operations * Monitor embodied carbon * Monitor & manage carbon in supply chain * Environmental monitoring & reporting * Optimize energy sourcing and supply chain |
| **Ecosystem / Real Estate** |  | * Reduce energy and electricity consumption by facilities and infrastructure * Green Building Ecosystem * Reduce environmental impact on natural ecosystems, including species biodiversity, land / agriculture etc. | * Monitoring building efficiency * Monitor forests, lands, species * Assess environment impact |

Sustainable cloud positions companies to deliver on new commitments: carbon reduction and responsible innovation. Companies have historically driven financial, security, and agility benefits though cloud, but sustainability is becoming an imperative.

|  |  |  |
| --- | --- | --- |
| **44%** | **|** | of CEOs in the United Nations Global Compact – Accenture Strategy CEO study on Sustainability see a net-zero future for their company in the next ten years. |
| **4.7X** | **|** | Between 2013-2019, companies with consistently high environmental, social and governance (ESG) performance enjoyed 4.7X higher operating margins and lower volatility than low ESG performers over the same period. |
| **30-40%** | **|** | Migrations to public cloud result in up to 30-40% total cost of ownership (TCO) savings. |

Drivers like greater workload flexibility, better server utilization rates, and more energy-efficient infrastructure all make public clouds more efficient than enterprise-owned data centers.

# Competitive Landscape

* + - **Salesforce**: <https://www.salesforce.com/products/sustainability-cloud/overview/>
    - **AWS**: <https://sustainability.aboutamazon.com/environment/the-cloud#:~:text=Sustainability%20in%20the%20Cloud%20Amazon%20Web%20Services%20%28AWS%29,100%25%20renewable%20energy%20usage%20for%20our%20global%20infrastructure>.
    - **Google**: <https://sustainability.google/>

# FAQs

#### Sustainability FAQs: [FAQ (sharepoint.com)](https://microsoft.sharepoint.com/sites/sustainability/sitepages/faq.aspx)