ISA 419: Data Driven Security

12 - A Very Brief Introduction to Predictive Analytics

Fadel M. Megahed

Endres Associate Professor
Department of Information Systems and Analytics
Farmer School of Business
Miami University
Email: fmegahed@miamioh.edu
Office Hours: Automated Scheduler for Office Hours

Fall 2022

Outline

What we covered in the past couple of weeks?

- **Introduction to Data Encoding**
- **The importance of data visualization**
- **The principles of data Visualization**
- **An overview of clustering techniques**

Learning Objectives for Today's Class

- **Define what do we mean by machine learning**
- **Explain the different types of learning**
- **Describe the different steps in using machine learning for predictive modeling applications**

Outline

Definition

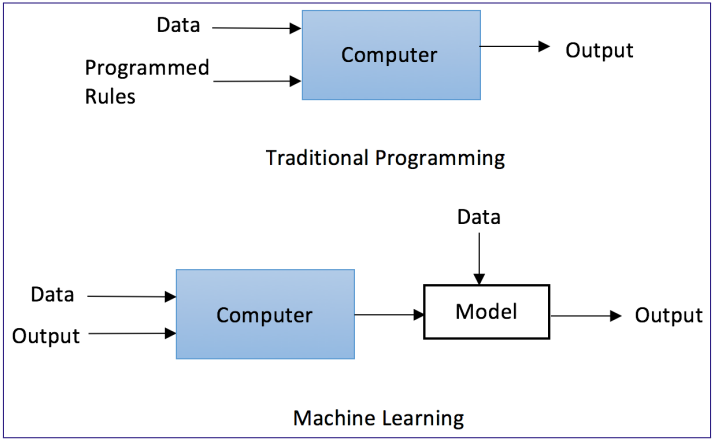
Mitchell (2006) has elgontely defined the scientific field of machine learning to be centered around answering the following question:

> *"How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?"*

In his view, machine learning is the study of algorithms that:

- **improve its performance** P
- **at task** T
- **following experience** E

A Paradigm Shift in Programming

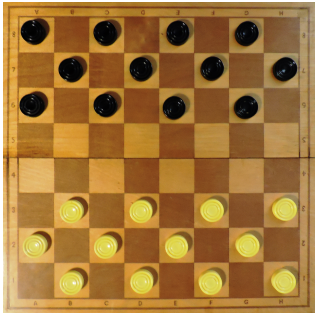


Source: Image is from Yuxi Lui (2019). Python Machine Learning by Example. Packt Publishers (click on image for more details).

Defining the Learning Task [1]

Improve on task T, with respect to performance metric P, based on experience E

- **T: Playing checkers**
- **P: Percentage of games won against an arbitrary opponent**
- **E: Playing practice games against itself**



Note: This idea in Samuel (1959) led to the popularization of machine learning.

Defining the Learning Task [2]

Improve on task T, with respect to performance metric P, based on experience E

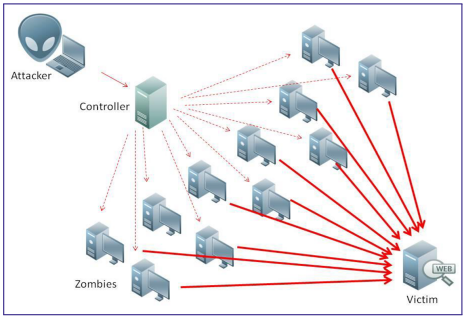
- **T: Autonomous driving using LADAR sensing**
- **P: Average distance traveled before human-judged error**
- **E: A sequence of images and steering commands recorded while observing a human driver**

Sources: Image by Dllu - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=64517567 and text from https://www.seas.upenn.edu/~cis519/fall2017/lectures/01_introduction.pdf

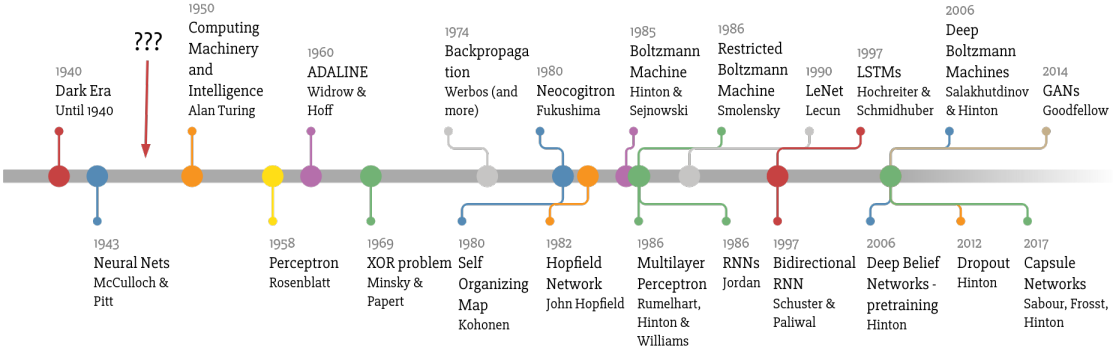
Defining the Learning Task [3]

Improve on task T, with respect to performance metric P, based on experience E

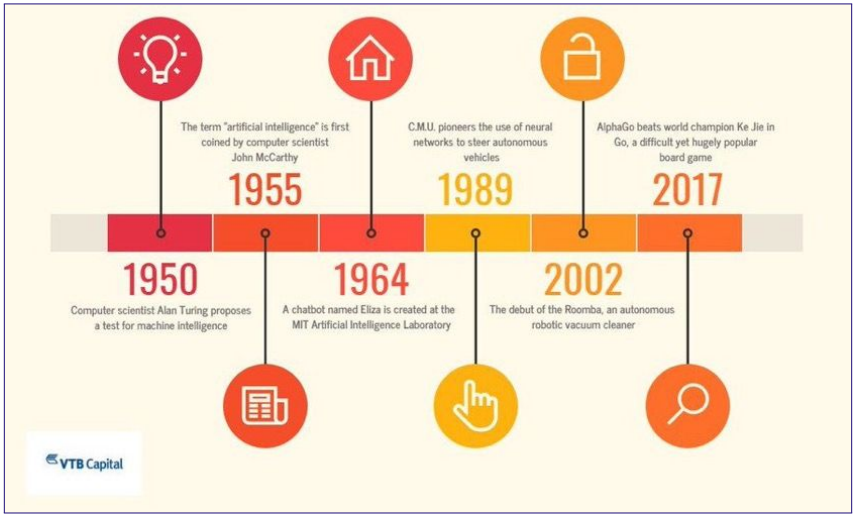
- **T: Categorizing network traffic as begnin or portmap (or another DDoS attack)**
- **P: Percentage of correctly categorized observations in each group**
- **E: Database of network traffic, with human given labels**



History of Machine Learning (click for source) [1]

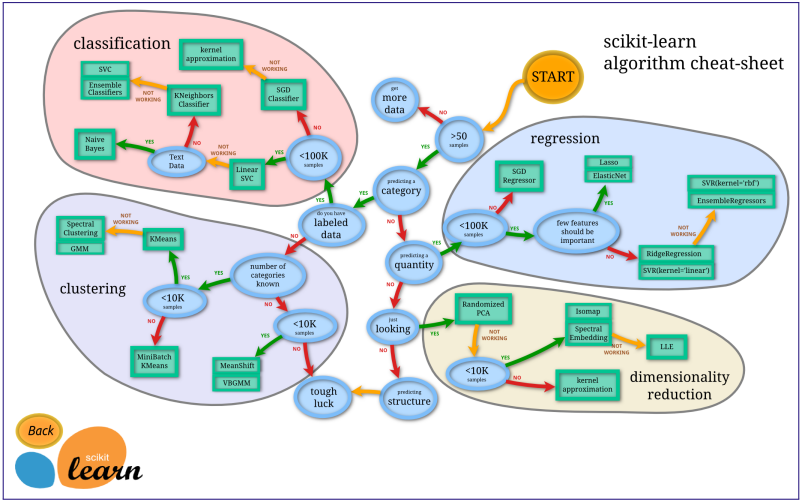


History of Machine Learning (click for source) [2]



Outline

Types of Learning (from Class 03)



Source: SciKit-Learn "Choosing the Right Estimator" (2020).

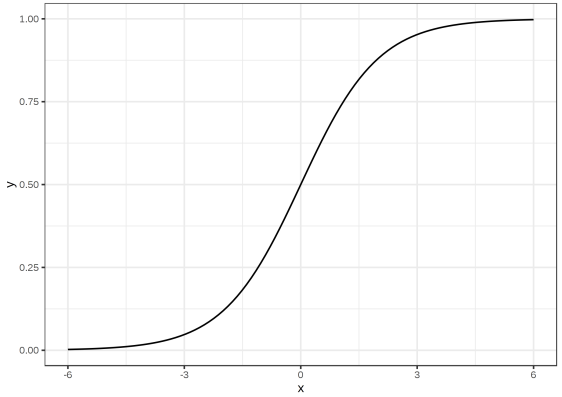
Supervised Learning: Regression [1]

- **Given:** (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n)
- **Learn a function $f(x)$ to predict y given x**
  - **If y is a real-valued variable (i.e. can be assumed to be continuous), then we do have a regression type problem.**
  - **Note in this class, we do not limit ourselves to simple and multiple linear regression. You should be open-minded to benchmarking their performance against other machine learning methodologies.**
    - *Simple linear regression:* single predictor $x \longrightarrow y$
    - *Multiple linear regression:* multiple predictors $X \longrightarrow y$; X is a vector

Supervised Learning: Regression [2]

```
# Build on the example below to answer the following questions:

# (A) Let us plot the data of x = ufo2010 and y = infections using ggplot()
#   (When we plot the data, let us fit a linear regression line in plot)

# (B) Let us examine how we can use R to create a simple linear regression
# on the same data -> summary(OBJECT)

# (C) Let us examine how we can use R to create a multiple linear regression
# on the same data -> summary(OBJECT)

# (D) "ISA 419 class links ZeroAccess Infections to Alien Visitors"??

# (E) What next?
```

Supervised Learning: Classification [1]



Source: Please click on image to access Ch 4.2 in Molnar (2019).

Supervised Learning: Classification [2]

- **Given:** (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n)
- **Learn a function $f(x)$ to predict y given x**
  - **If y is categorical (special case being binary i.e. 2 categories), then we have a classification problem. As in the regression case, the statistical model logistic regression (and/or one of its varients such as LASSO, Ridge Regression, etc) are often used as benchmark model(s)**
  - **For classification problems, let assume that we have a binary classification problem where the outcome y is defined as**

Supervised Learning: Classification [3]

$$y = \begin{cases} 0, & \text{if no attack is happening/successful} \\ 1, & \text{otherwise} \end{cases} \tag{1}$$

In such a case, the probability of an attack (i.e. $y = 1$) can be defined as

$$P(y = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p))}. \tag{2}$$

To ease the interpretation, we can reformulate the equation above as

$$log\left(\frac{P(y = 1)}{1 - P(y = 1)}\right) = log\left(\frac{P(y = 1)}{P(y = 0)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p. \tag{3}$$

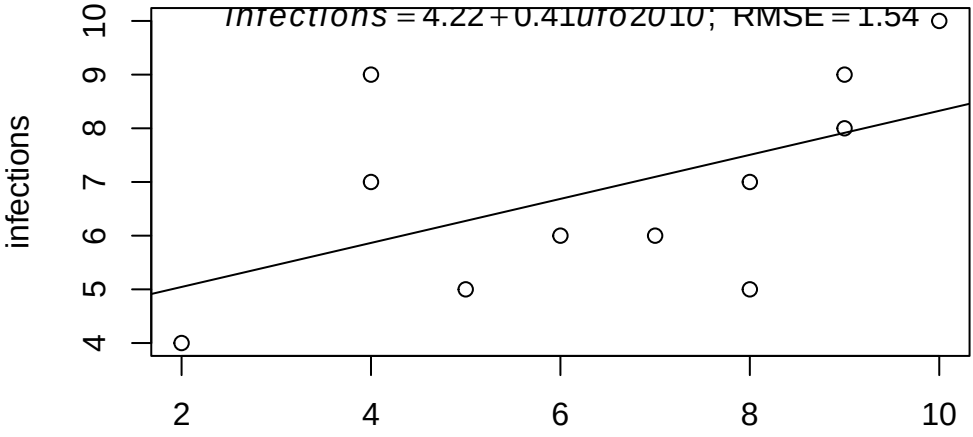
Supervised Learning: Classification [4]

See Sec. 4.2 in Molnar (2019) for details on model interpretation .

Outline

Performance Metrics in Regression Problems

In ML regression problems, the root mean squared error (RMSE) is typically used. It measures the standard deviation of the residuals (prediction errors).



Performance Metrics in Classification Problems

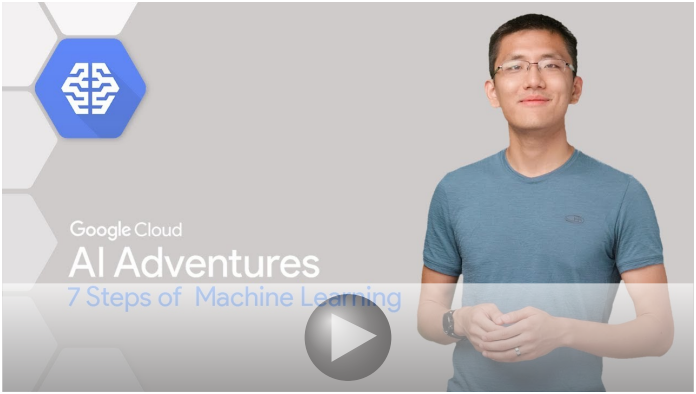
In binary classification problems, most metrics can be captured from the confusion matrix. Below, I include a synthatic example for a confusion matrix for a binary classifier, which we will use to define the following terms:

		Actual	
		Event	**No Event**
Pred.	**Event**	A	B
	No Event	C	D

- Accuracy
- Sensitivity
- Specificity
- Balanced Accuracy
- Precision
- Recall

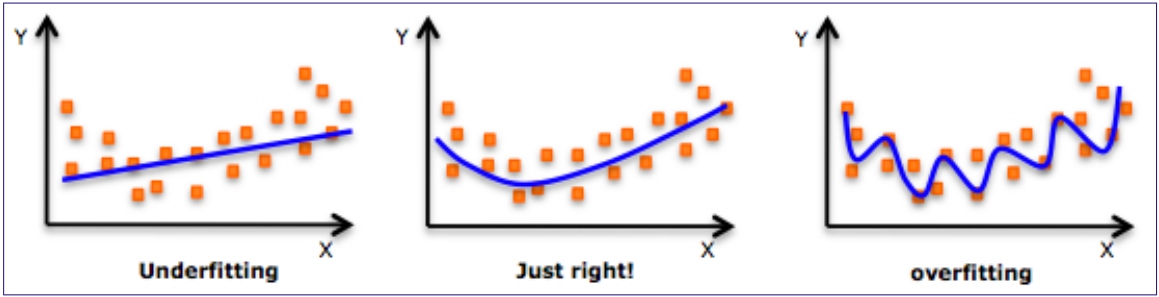
Outline

The 7 Steps of ML



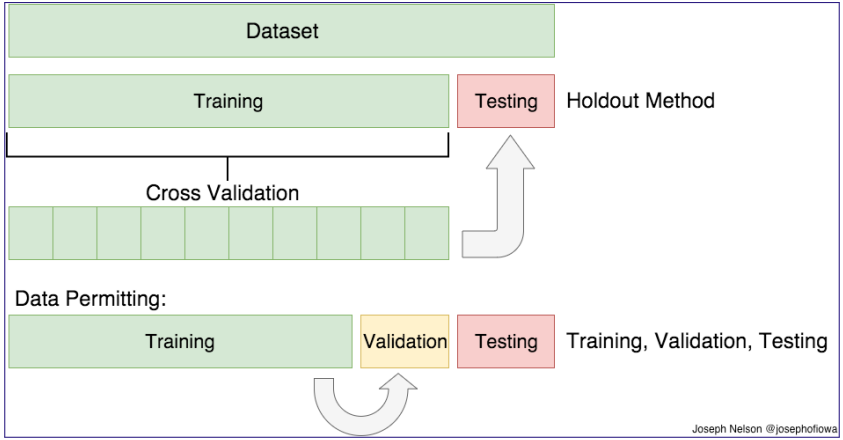
For an overview of these steps, please refer to the following medium article.

Comments: Preparing the Data for Modeling [1]



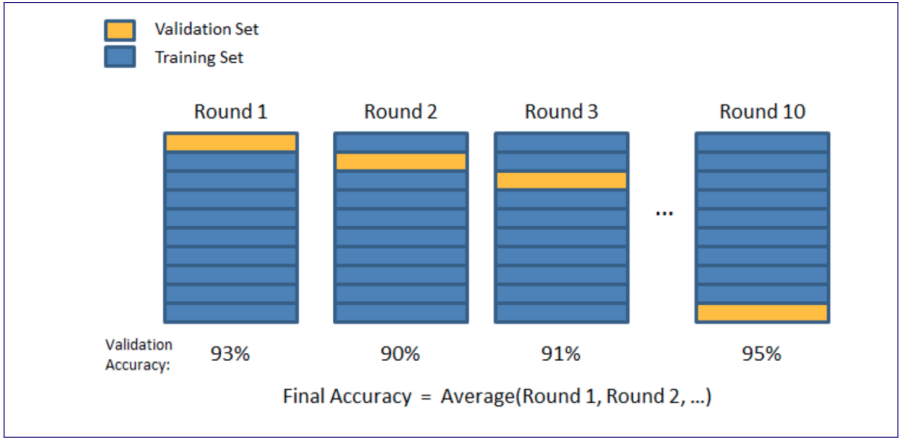
Source: Adi Bronshtein (2017); Towards Data Science.

Comments: Preparing the Data for Modeling [2]



Source: Adi Bronshtein (2017); Towards Data Science.

Comments: Preparing the Data for Modeling [3]



Source: Adi Bronshtein (2017); Towards Data Science.

Comments: Need for Resampling in Unbalanced Data [1]

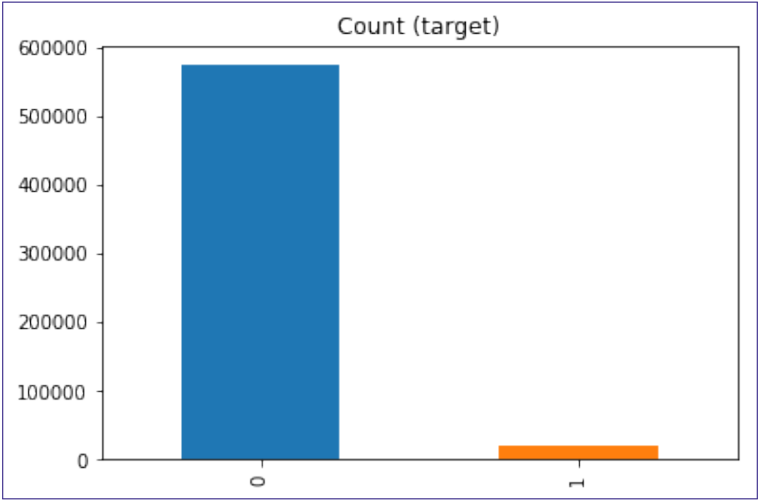
In many security datasets, actual events (e.g., breaches will be rare). When training your model, you will need to account for the rareness by attempting to balance the training data (so that you do not end up with a naive model).
Typical re-sampling techniques include:

- Random under-sampling;
- Random over-sampling;
- Synethetic majority oversampling technique;

Comments: Need for Resampling in Unbalanced Data [2]

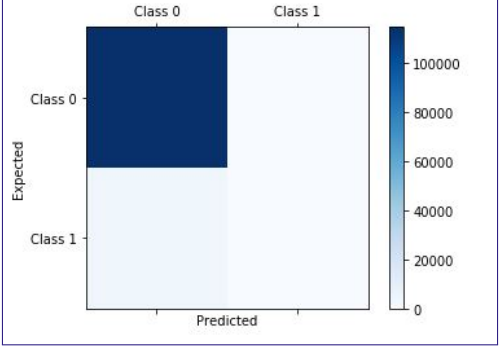
The images in the following slides are based on an example by Rafael Alencar.

Comments: Need for Resampling in Unbalanced Data [3]

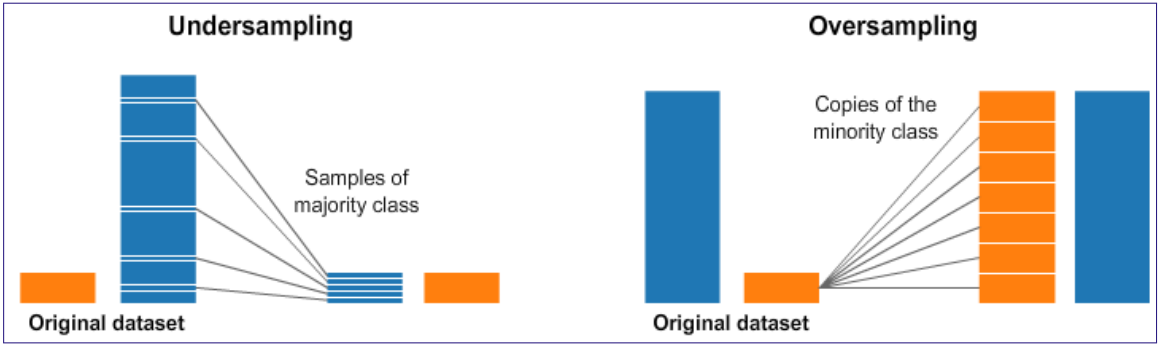


Comments: Need for Resampling in Unbalanced Data [4]

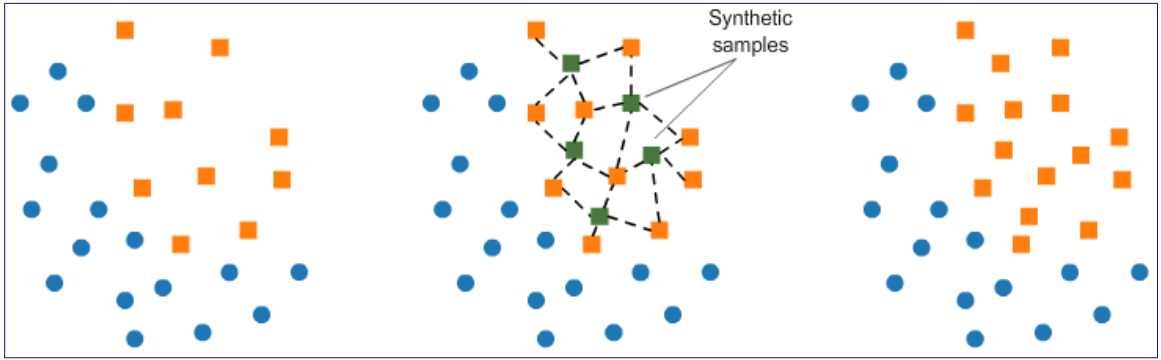
```
Confusion matrix:
 [[114709      0]
 [  4334      0]]
```



Comments: Need for Resampling in Unbalanced Data [5]



Comments: Need for Resampling in Unbalanced Data [6]



Outline

Today's Learning Objectives

- **Define what do we mean by machine learning**
- **Explain the different types of learning**
- **Describe the different steps in using machine learning for predictive modeling applications**

References [1]

"Choosing the Right Estimator." 2020. https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html.

Mitchell, Tom Michael. 2006. "The Discipline of Machine Learning." Machine Learning Department. School of Computer Science, Carnegie Mellon University. http://ra.adm.cs.cmu.edu/anon/usr0/ftp/anon/ml/CMU-ML-06-108.pdf.

Molnar, Christoph. 2019. "Interpretable Machine Learning." Lulu.com. https://christophm.github.io/interpretable-ml-book/.

Samuel, Arthur L. 1959. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of Research and Development* 3 (3): 210–29. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5392560.

ISA 419: Data Driven Security

12 - A Very Brief Introduction to Predictive Analytics

Fadel M. Megahed

Endres Associate Professor
Department of Information Systems and Analytics
Farmer School of Business
Miami University
Email: fmegahed@miamioh.edu
Office Hours: Automated Scheduler for Office Hours

Fall 2022