

# NDL GBM #3

## Bootcamp Day 2

### Modeling

9/21/25

# Attendance: Meeting Name “Modeling”





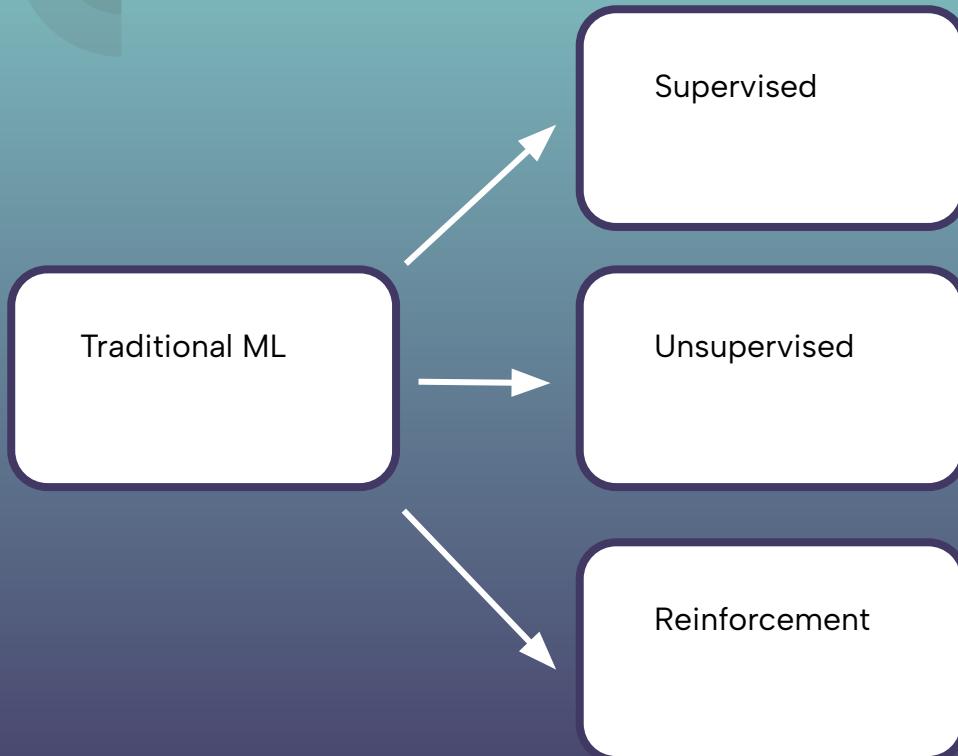
## What is ML?

Traditional ML: Using the known to predict the unknown by training an algorithm

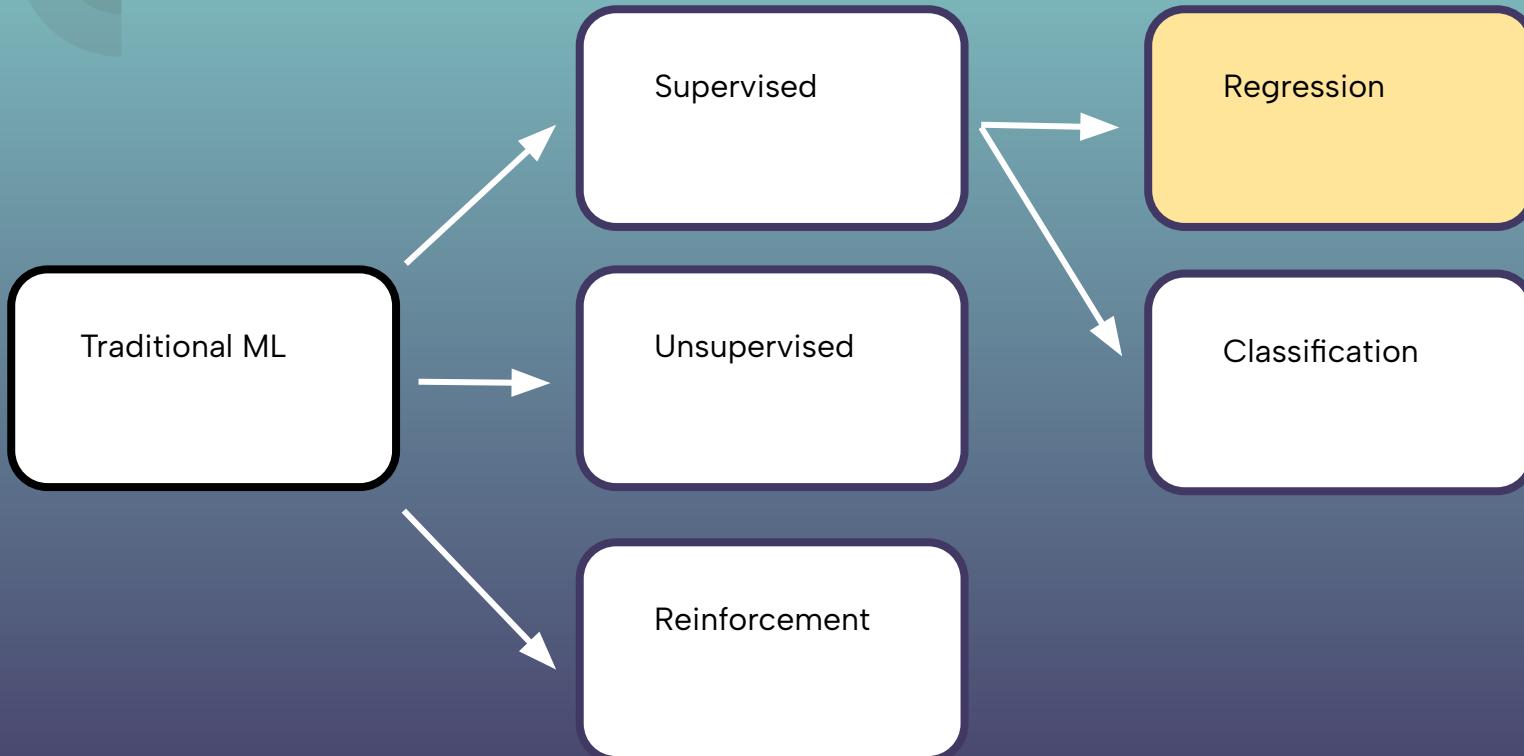
## What kind can achieve our goal?

Regression - predicting quantitative variables

# Where does regression fit?

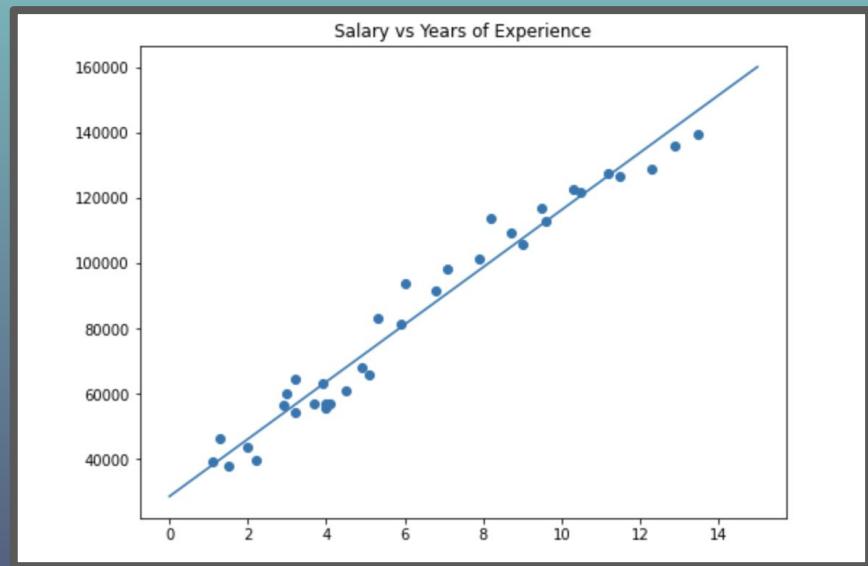


# Where does regression fit?



# Simple linear regression

$$y = mx + b$$



$$\text{Salary} = 1000(\text{Y.o.E}) + 1000$$

# Getting m and b

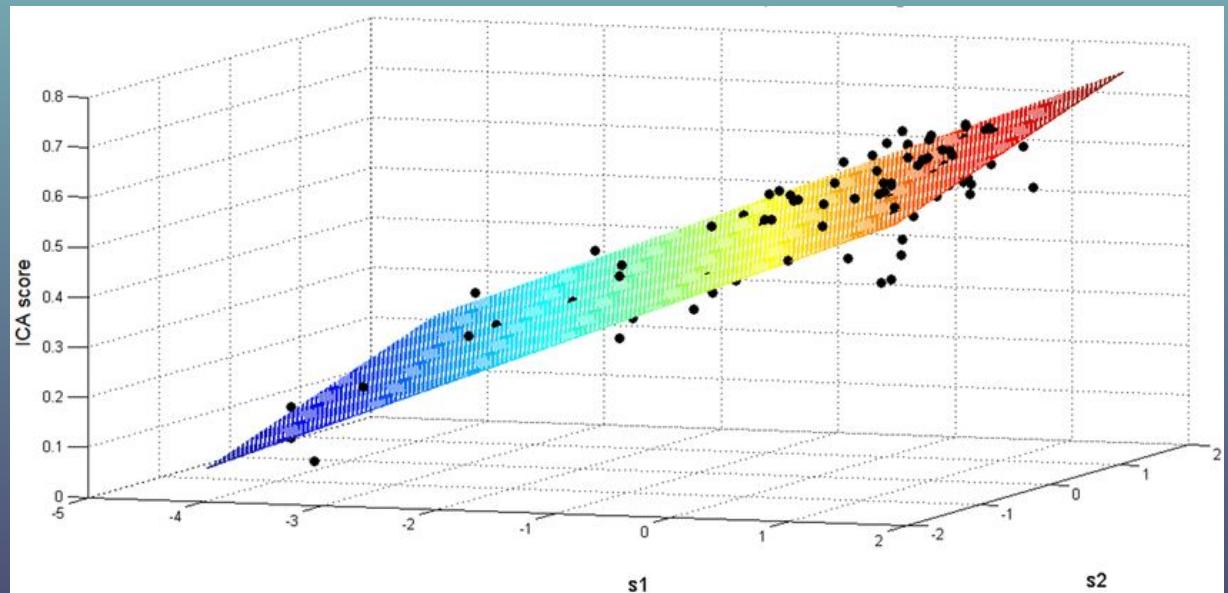
$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}^1 \\ \hat{y}^2 \\ \hat{y}^3 \\ \hat{y}^4 \\ \hat{y}^5 \\ \hat{y}^6 \\ \hat{y}^7 \\ \hat{y}^8 \\ \hat{y}^9 \\ \hat{y}^{10} \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{living}^1 \\ 1 & x_{living}^2 \\ 1 & x_{living}^3 \\ 1 & x_{living}^4 \\ 1 & x_{living}^5 \\ 1 & x_{living}^6 \\ 1 & x_{living}^7 \\ 1 & x_{living}^8 \\ 1 & x_{living}^9 \\ 1 & x_{living}^{10} \end{bmatrix}$$
$$\mathbf{w} = \begin{bmatrix} b \\ w \end{bmatrix} \longrightarrow \hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$$
$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$$
$$\mathbf{X}^T\hat{\mathbf{y}} = \mathbf{X}^T\mathbf{X}\mathbf{w}$$

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{y}}$$



$$y = m_1 x_1 + \\ m_2 x_2 + b$$

# Multi-Linear Regression



# Multi-linear regression

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}^1 \\ \hat{y}^2 \\ \hat{y}^3 \\ \hat{y}^4 \\ \hat{y}^5 \\ \hat{y}^6 \\ \hat{y}^7 \\ \hat{y}^8 \\ \hat{y}^9 \\ \hat{y}^{10} \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} x_0^1 = 1 & x_1^1 & x_2^1 \\ x_0^2 = 1 & x_1^2 & x_2^2 \\ x_0^3 = 1 & x_1^3 & x_2^3 \\ x_0^4 = 1 & x_1^4 & x_2^4 \\ x_0^5 = 1 & x_1^5 & x_2^5 \\ x_0^6 = 1 & x_1^6 & x_2^6 \\ x_0^7 = 1 & x_1^7 & x_2^7 \\ x_0^8 = 1 & x_1^8 & x_2^8 \\ x_0^9 = 1 & x_1^9 & x_2^9 \\ x_0^{10} = 1 & x_1^{10} & x_2^{10} \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$$

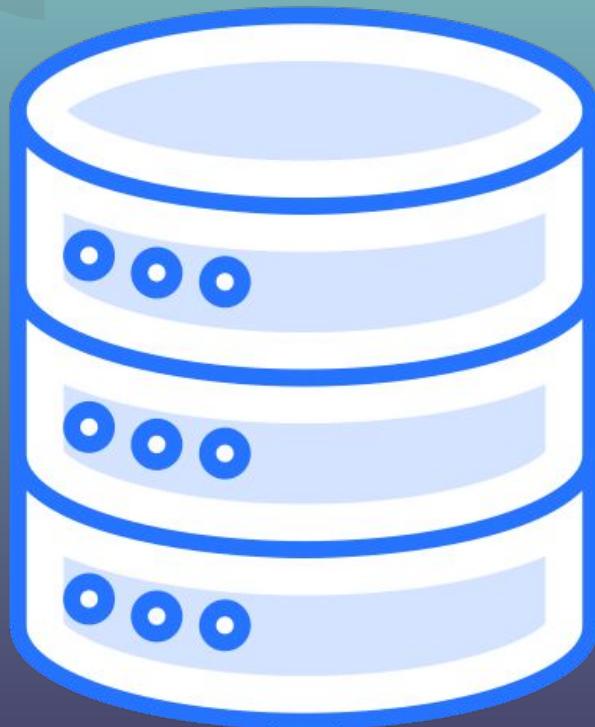
$\longrightarrow \hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$

$\downarrow$

$\mathbf{X}^T \hat{\mathbf{y}} = \mathbf{X}^T \mathbf{X}\mathbf{w}$

$\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{y}}$

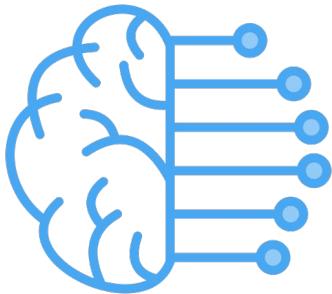
# Data Split



- Train – fit the model
- Validation – optimize the model
- Test – Evaluate the model

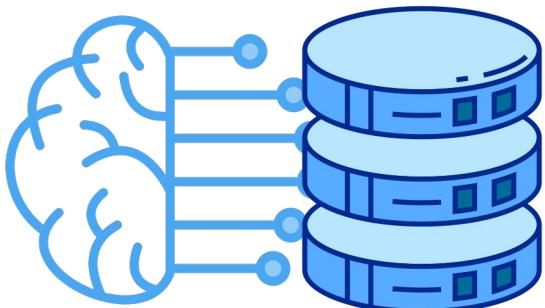


# How's it done?



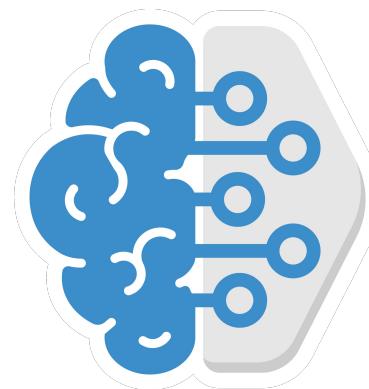
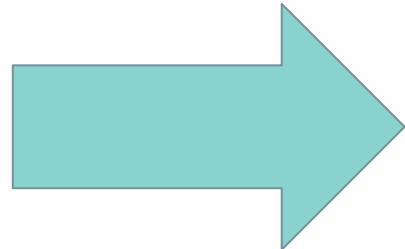
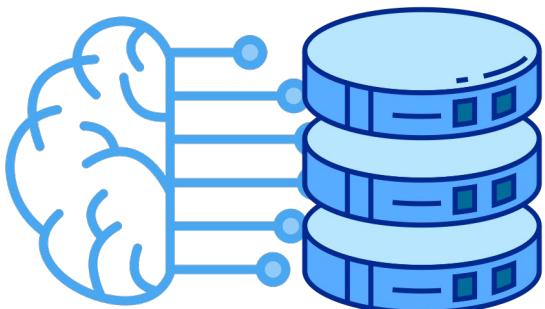


# How's it done?



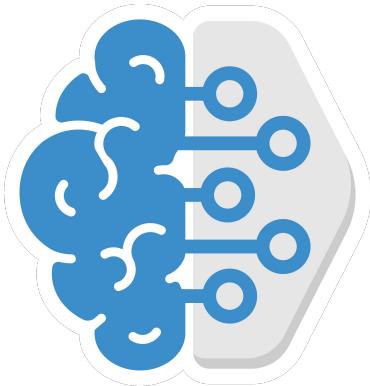


# How's it done?



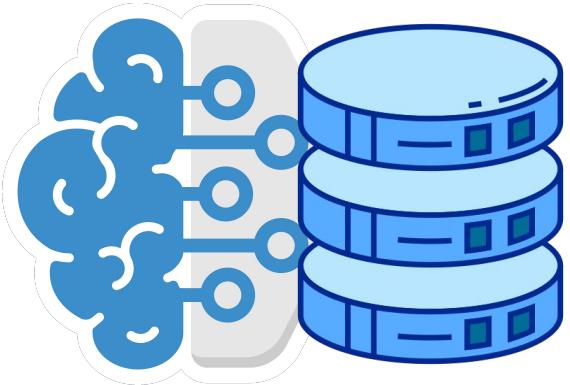


# How's it done?



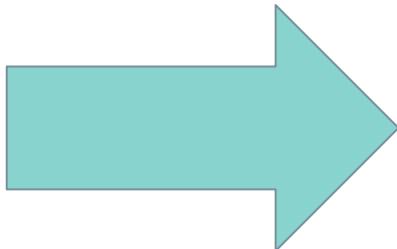
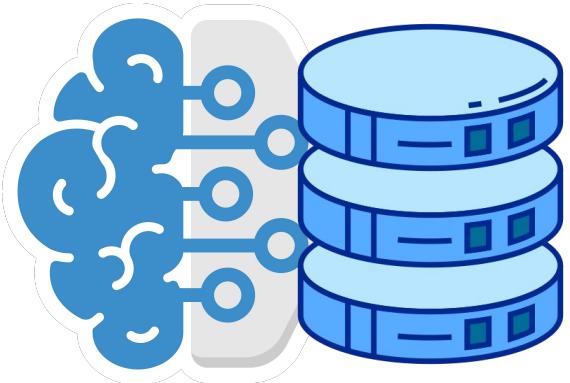


# How's it done?





# How's it done?



## How effective was it?

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Better interpretability

Punishes high errors more

$$MSE = \frac{1}{n} \sum (y - \hat{y})^2$$



## Example of difference in MAE/MSE

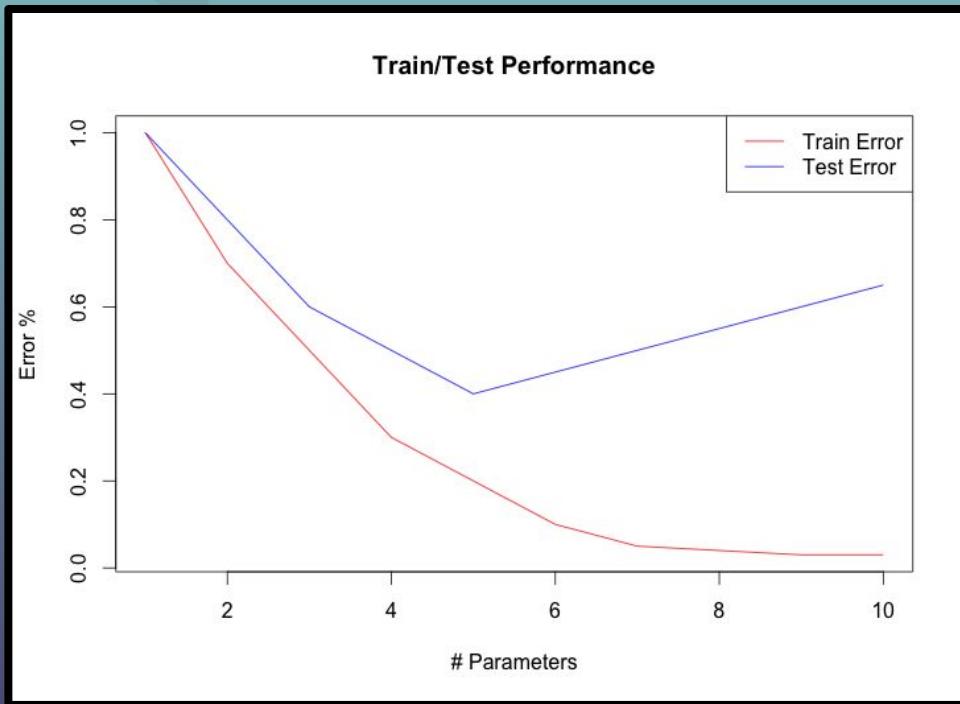
### Situation 1:

Pred = 0.5  
Actual = 0.7  
MAE = 0.2  
MSE = 0.16

### Situation 2:

Pred = 5  
Actual = 7  
MAE = 2  
MSE = 4

# Overfitting



Model becomes too used to the training error that the testing error increases



# Competition Details

- \$100 to winner, \$50 to 2nd, \$25 to 3rd
- ~a week and a half
- Requirements for certificate:
  - Come to all three bootcamp days **OR**
  - Come to at least one day and get a error of less than 449.4 in the competition

# Kaggle Submissions

The screenshot shows the Kaggle website interface for the "Airbnb NYC Price Prediction" competition. The left sidebar includes links for Create, Home, Competitions, Datasets, Models, Benchmarks, Code, Discussions, Learn, and More. Under "Your Work", there is a "VIEWED" section. The main content area displays the competition details:

- Competition Host:** Adam Torres Encarnacion
- Prizes & Awards:** Kudos (Does not award Points or Medals)
- Participation:** 1 Entry, 0 Participants, 0 Teams, 0 Submissions
- Tags:** Root Mean Squared Error
- Description:** Predict the price (nightly rental rate in USD) for Airbnb listings in the test set. You'll need to:
  - Explore and Clean the Data: Handle missing values, outliers, and data inconsistencies
  - Engineer Meaningful Features: Extract insights from location, host behavior, and property characteristics
  - Build Predictive Models: Develop machine learning algorithms that generalize well
- Table of Contents:** Overview, Description, Evaluation

A large red arrow points to the "Submit Prediction" button in the top right corner of the main content area.

# Kaggle Submissions

The screenshot shows the Kaggle web interface. On the left, the main navigation bar includes 'kaggle' (with a dropdown menu), a search bar, and various navigation links like Home, Competitions, Datasets, Models, Benchmarks, Code, Discussions, Learn, and More. A sidebar on the far left lists 'Your Work' and 'Viewed'. The central area displays a competition titled 'Airbnb NYC Price Prediction' under the 'NYC Airbnb Price Prediction Challenge'. The competition details include an overview, competition overview, start date (2 days ago), and a description section. To the right, a modal window titled 'Submit to Competition' is open for the 'Airbnb NYC Price Prediction' competition. The modal has tabs for 'File Upload' (selected) and 'Notebook'. It features a large file upload area with a placeholder 'Drag and drop file to upload (e.g., .csv, .zip, .gz, .7z)' and a 'Browse Files' button. Below this is a 'SUBMISSION DESCRIPTION' field with a placeholder 'Enter a description'. At the bottom, there's a terminal-like input field with the command '> \_ kaggle competitions submit -c airbnb-nyc-price-prediction -f\_ ...', a 'Cancel' button, and a prominent 'Submit' button.

# Kaggle Submissions

The screenshot shows a web browser displaying the Kaggle website. On the left, the main navigation menu is visible, including 'Create', 'Home', 'Competitions', 'Datasets', 'Models', 'Benchmarks', 'Code', 'Discussions', 'Learn', and 'More'. Below this is a sidebar for 'Your Work' with a 'VIEWED' section. The main content area is titled 'Airbnb NYC Price Prediction' and includes tabs for 'Overview', 'Data', 'Code', 'Models', 'Discussion', 'Leaderboard', and 'Rules'. The 'Overview' tab is selected. It features a 'NYC Airbnb Price Prediction Challenge' section with a brief description and a timeline starting '2 days ago'. Below this is a 'Description' section with a 'Your Task' paragraph and a numbered list of three steps: 'Explore and Clean the Data', 'Engineer Meaningful Features', and 'Build Predictive Models'. A large modal window titled 'Submit to Competition' is overlaid on the page. This modal has tabs for 'File Upload' (which is selected) and 'Notebook'. It displays the competition title 'Airbnb NYC Price Prediction' and a message indicating 'You have 2 submissions remaining today. This resets in 7 hours.' It contains a large input field with a placeholder 'Drag and drop file to upload (e.g., .csv, .zip, .gz, .7z)' and a 'Browse Files' button. Below this is a 'SUBMISSION DESCRIPTION' field with the placeholder 'Enter a description'. At the bottom of the modal is a 'SUBMIT USING THE KAGGLE API' section containing a command-line interface (CLI) command: `>_ kaggle competitions submit -c airbnb-nyc-price-prediction -f...`. Finally, there are 'Cancel' and 'Submit' buttons at the bottom right of the modal.



# Epoch: Your Research Reading Club

Passionate about delving into the latest research? Epoch is a community where curious minds come together to explore, discuss, and learn.

## What We Offer:

- In-depth modern paper discussions.
- Opportunities to expand your knowledge.
- A chance to present and prove yourself to your community!

Epoch provides a supportive environment for intellectual growth.

## Join Our Community!

Scan the QR code to learn more and become a member.





# Next Meeting

Will be in Westgate e203

Recap Lab

Help session