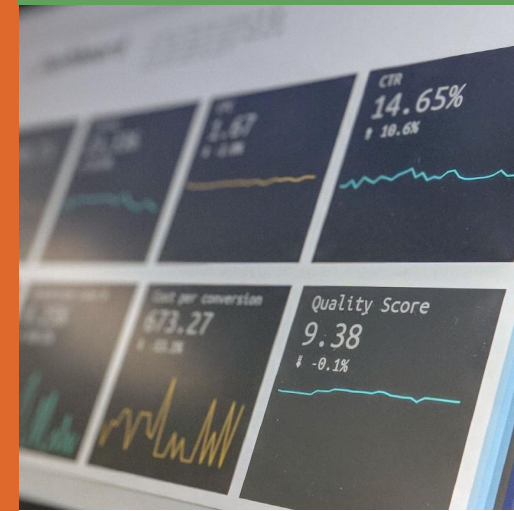
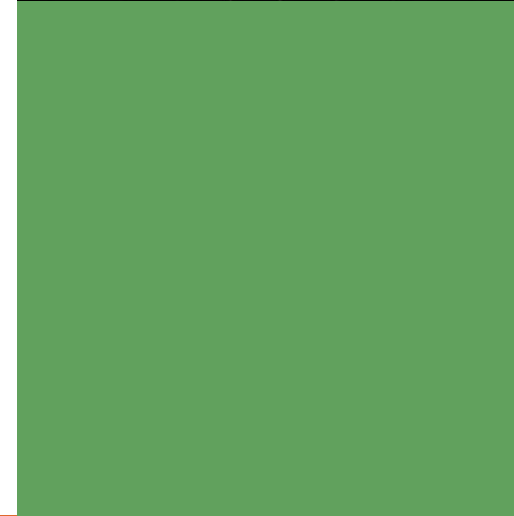


RESPONSIBLE ARTIFICIAL INTELLIGENCE LAB (RAIL)

Machine Learning with Regression

6th December 2022



OUTLINE

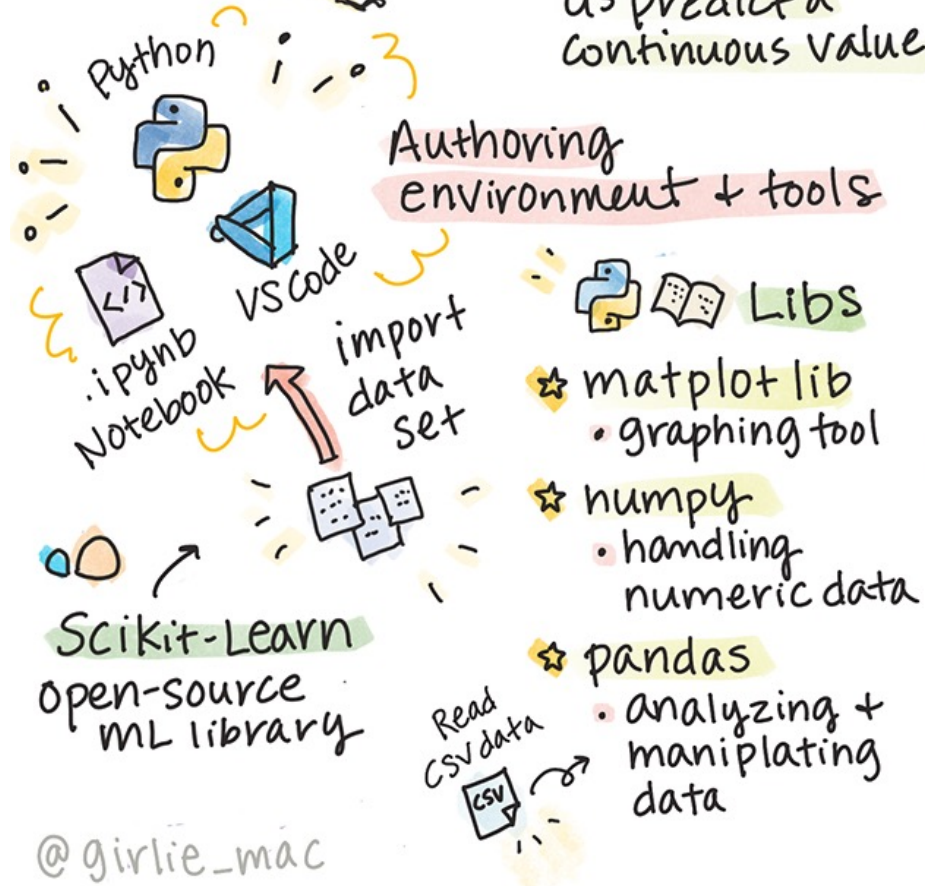


1. Configure your computer for local machine-learning tasks
2. Work with Jupyter notebooks
3. Use Scikit-learn, including installation
4. Explore linear regression with a hands-on exercise
5. Explore Logistic Regression with hands-on exercise

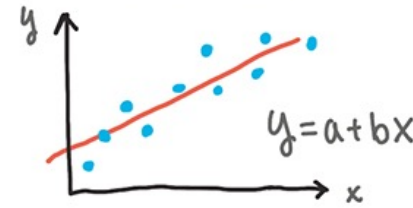
Machine Learning Regression

@Azure Advocates

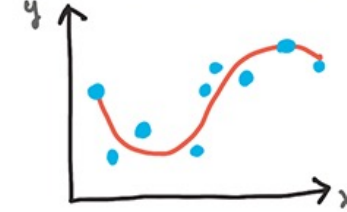
mathematical methods that let us predict a continuous value



Linear Regression



Polynomial Regression

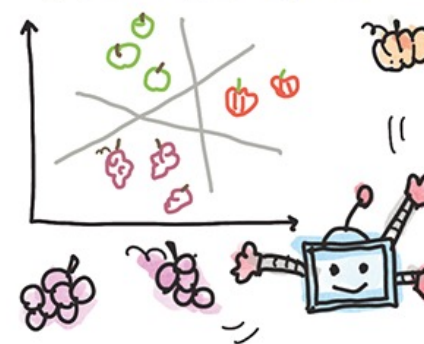


Logistic Regression

Binary Classification



Multinomial Classification



Ordinal Classification



Installations and Configurations

- Steps
 - Visit <https://scikit-learn.org/stable/install.html> and follow the instructions to install scikit-learn. It also has instructions for Apple M1.
 - For pip, use the command “pip install scikit-learn”
 - For conda, use the command “conda install scikit-learn”



Machine Learning with Scikit-Learn

Your ML Environment

- You will use notebooks to develop your Python code and create machine-learning models.
- This file type is a standard tool for data scientists, and they can be identified by their suffix or extension `.ipynb`.

Notebooks are an interactive environment allowing developers to code and add notes.

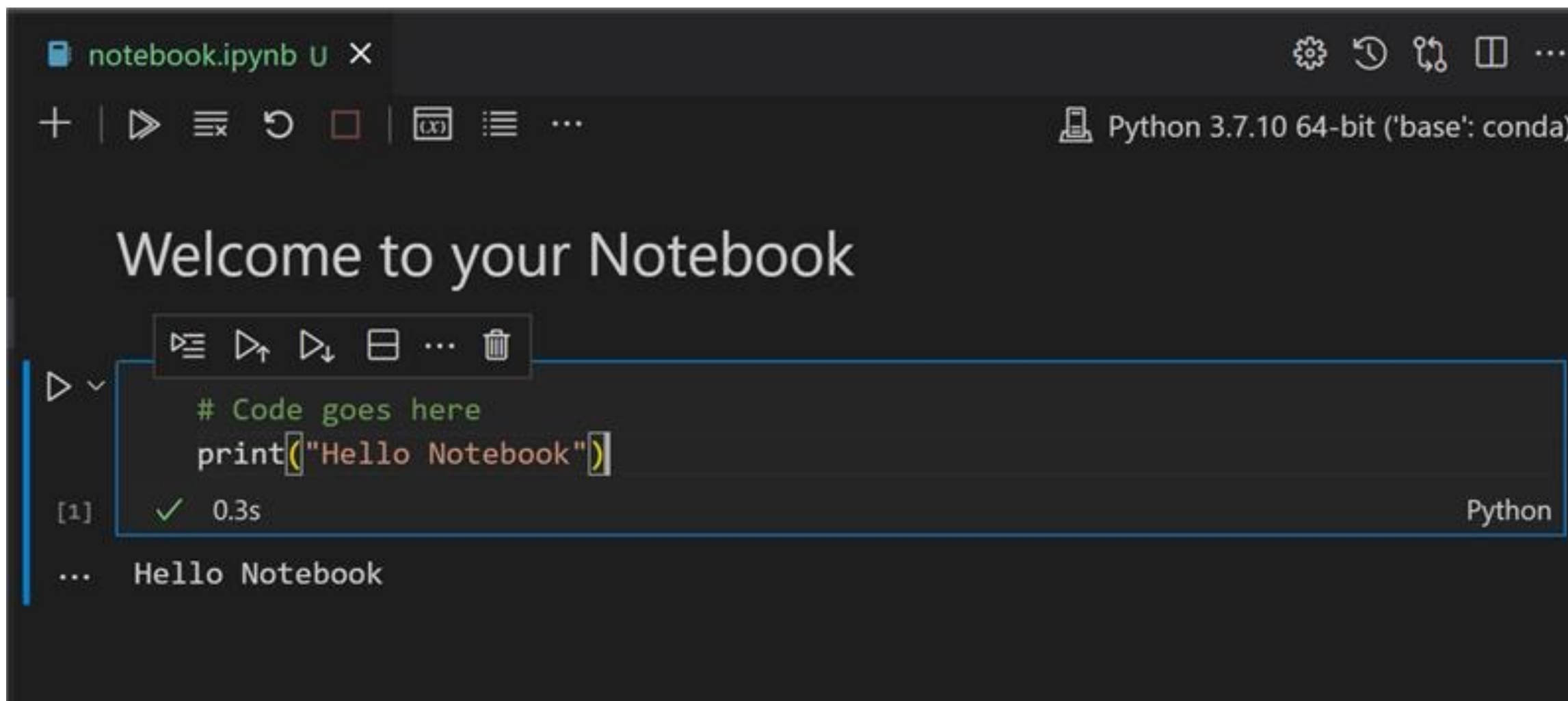
It also allows developers to write documentation around the code which is quite helpful for experimental or research-oriented projects.

Activity – Work with a Notebook

- Create a file "notebook.ipynb" in your base folder from Visual Studio Code.
- A Jupyter server will start with Python 3+ started.
- You can run a code block by selecting the icon that looks like a play button.
- Select the markdown icon and add a bit of markdown and the text "# Welcome to your notebook".
- Next, add some Python code.
- Type `print('hello notebook')` in the code block.
- Select the arrow to run the code.



Activity – Work with a Notebook



The screenshot shows a Jupyter Notebook window titled 'notebook.ipynb'. The interface includes a top toolbar with icons for file operations, a code editor with a 'Welcome to your Notebook' message, and a cell containing the code `print("Hello Notebook")`. The code has been executed, resulting in the output 'Hello Notebook'. The environment is identified as 'Python 3.7.10 64-bit ('base': conda)'.

```
# Code goes here
print("Hello Notebook")
```

[1] ✓ 0.3s Python

... Hello Notebook

Up and Running with Scikit-learn

- Now that Python is set up in your local environment and you are comfortable with Jupyter notebooks let's get equally comfortable with Scikit-learn (pronounce it sci as in science).
- Scikit-learn provides an extensive API to help you perform ML tasks.
- Scikit-learn is an open-source machine-learning library that supports supervised and unsupervised learning.
- It also provides tools for model fitting, data pre-processing, model selection and evaluation, and many other utilities.

Your First Scikit-Learn Notebook

- You will work with a small dataset about diabetes that is built into Scikit-learn for learning purposes.
- Imagine that you wanted to test a treatment for diabetic patients.
- .
- ML models might help you determine which patients would respond better to the treatment based on combinations of variables.
- When visualised, even a very basic regression model might show information about variables that would help you organise your theoretical clinical trials.

Your First Scikit-Learn Notebook

- There are many types of regression methods, and which one you pick depends on the answer you're looking for.
- If you want to predict the probable height for a person of a given age, you'd use linear regression, as you're seeking a numeric value.
- If you're interested in discovering whether a type of cuisine should be considered vegan or not, you're looking for a category assignment so you would use logistic regression.
- Think a bit about some questions you can ask of data, and which of these methods would be more appropriate.

Activity: First Scikit-Learn Notebook

- Go through the Jupyter notebook on the First Scikit-Learn Notebook.
- Click on Image to Open the File
- Plot a different variable from this dataset.



Edit this line: $X = X[:, np.newaxis, 2]$. Given this dataset's target, what can you discover about the progression of diabetes as a disease?

Knowledge Check - 1

- If your Machine Learning model's training accuracy is 95% and the testing accuracy is 30%, then what type of condition is it called?

A. Overfitting

B. Underfitting

C. Double fitting

Knowledge Check - 2

- The process of identifying significant features from a set of features is called?

A. Feature Extraction

B. Feature Dimensionality Reduction

C. Feature Selection

Knowledge Check - 3

- The process of Splitting a dataset into a specific ratio of training and testing datasets using Scikit-Learn's “train_test_split()” method/function is called?

A. Cross-Validation

B. Hold-out Validation

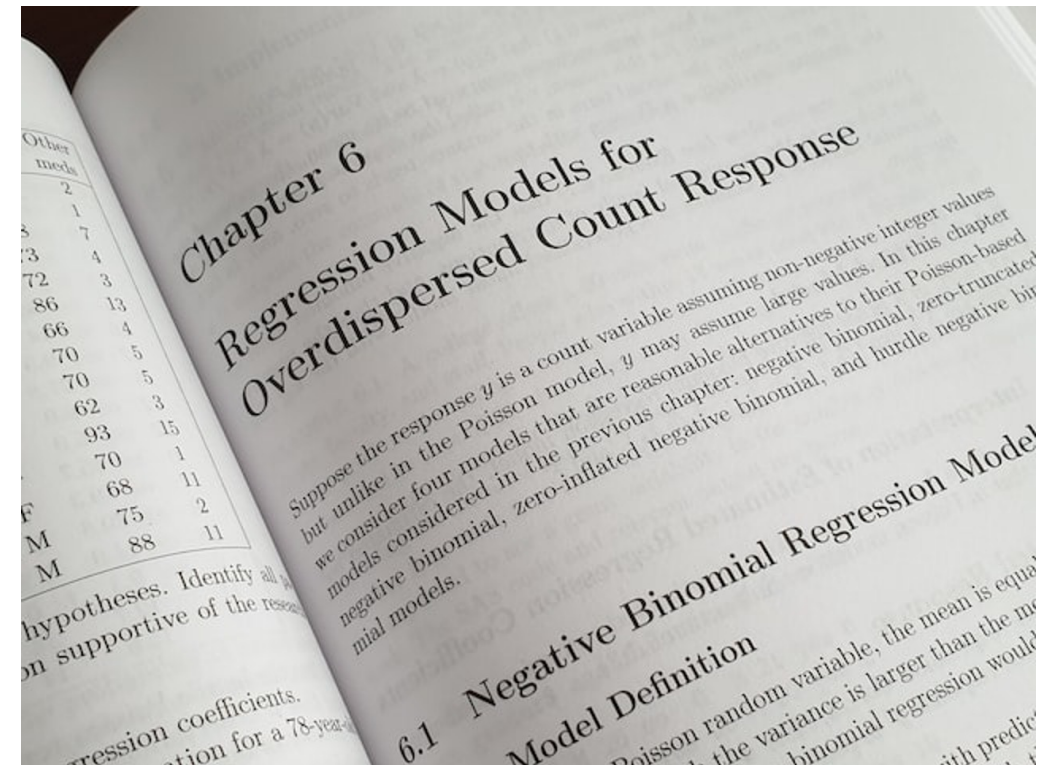
C. Leave one out Validation

What is Regression

- Regression works by establishing a relationship between variables in the data that represent characteristics—known as the features—of the observed thing and the variable we're trying to predict—known as the label.
- For example, to predict the expected number of rentals in a given day by a bike company.
- To train the model, we start with a data sample containing the features and known values for the label.
- So, in this case, we need historical data that includes **dates, weather conditions,** and the **number of bicycle rentals.**

Activity: Train and Evaluate a Regression Model

- Go through the Jupyter notebook on the Train and Evaluate a Regression Model.
- Click on Image to Open the File

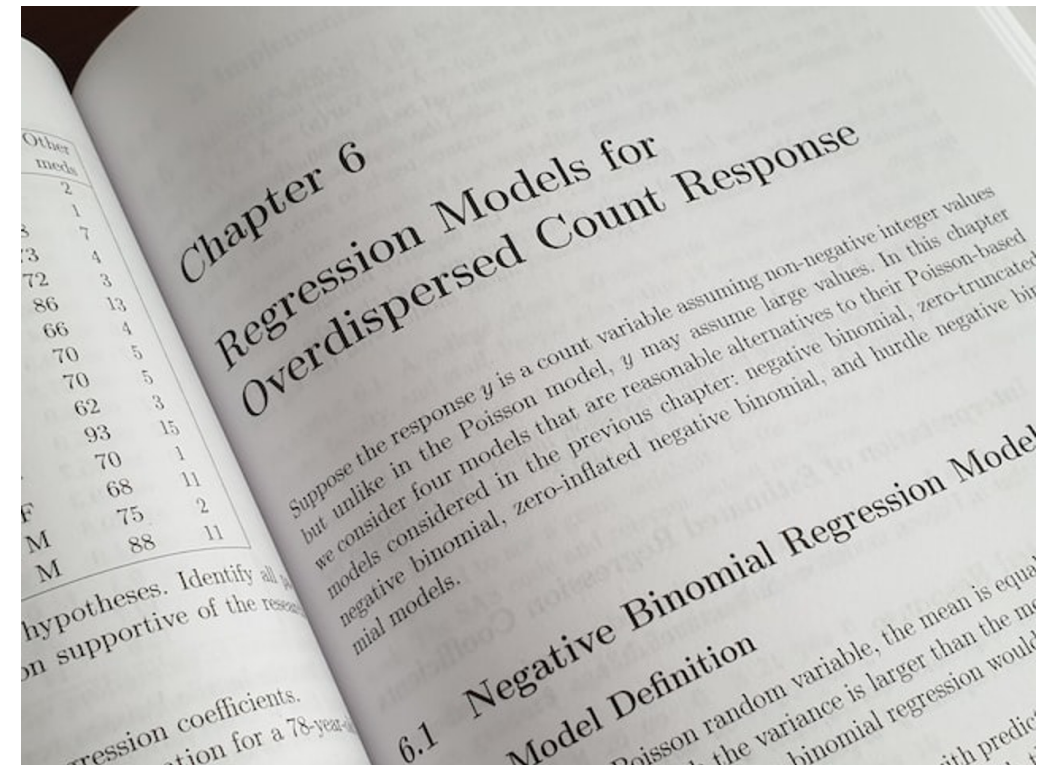


Discover New Regression Models

- Regression models are often chosen because they work with small data samples, are robust, are easy to interpret, and exist in various forms.
- The various forms include:
 - Linear Regression
 - Decision Trees
 - Ensemble algorithms
- Data scientists often experiment using different models.
- In the next activity, we'll experiment with different models to compare their performance on the same data.

Activity: Experiment with More Powerful Regression Models

- Go through the Jupyter notebook on the Experiment with more Powerful Regression Models.
- Click on Image to Open the File

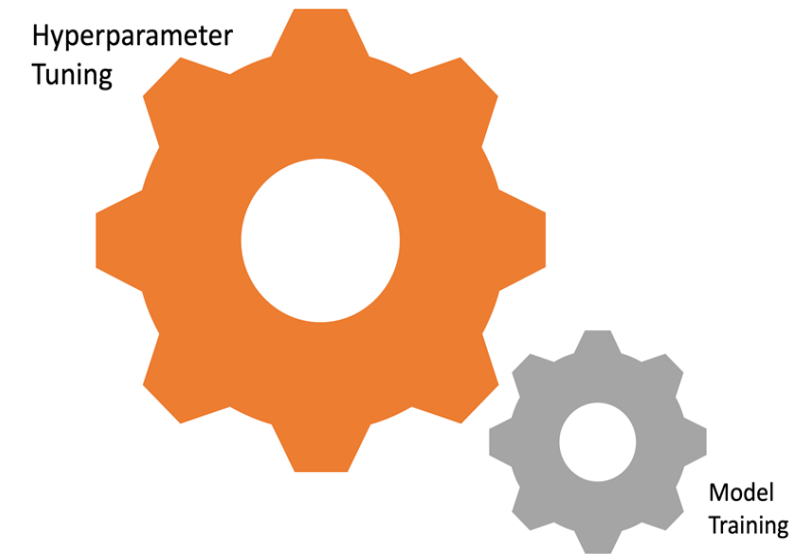


Improve Models with Hyperparameters

- Simple models with small datasets can often be fit in a single step, while more extensive datasets and more complex models must be fit by repeatedly using the model with training data and comparing the output with the expected label.
- If the prediction is accurate enough, we consider the model trained. If not, we adjust the model slightly and loop again.
- Hyperparameters are values that change how the model is fit during these loops.

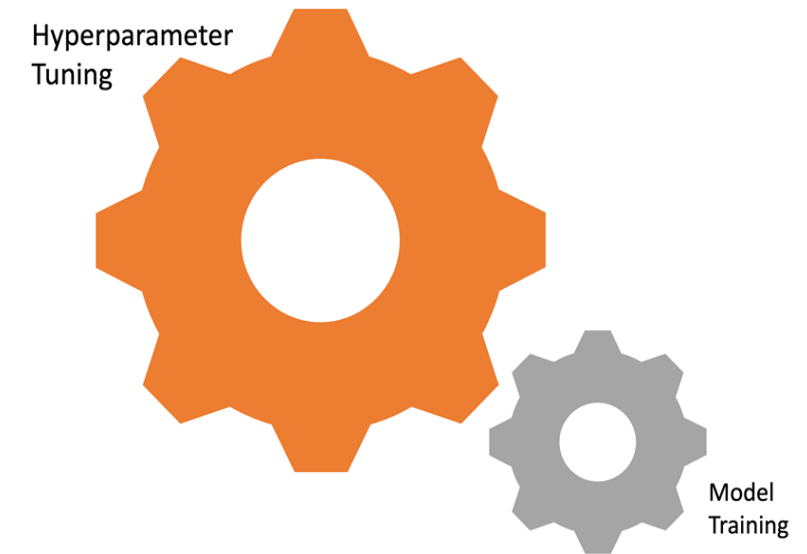
Improve Models with Hyperparameters

- Learning rate, for example, is a hyperparameter that sets how much a model is adjusted during each training cycle.
- A high learning rate means a model can be trained faster, but if it's too high, the adjustments can be so large that the model is never 'finely tuned' and is not optimal.



Improve Models with Hyperparameters

- Learning rate, for example, is a hyperparameter that sets how much a model is adjusted during each training cycle.
- A high learning rate means a model can be trained faster, but if it's too high, the adjustments can be so large that the model is never 'finely tuned' and is not optimal.

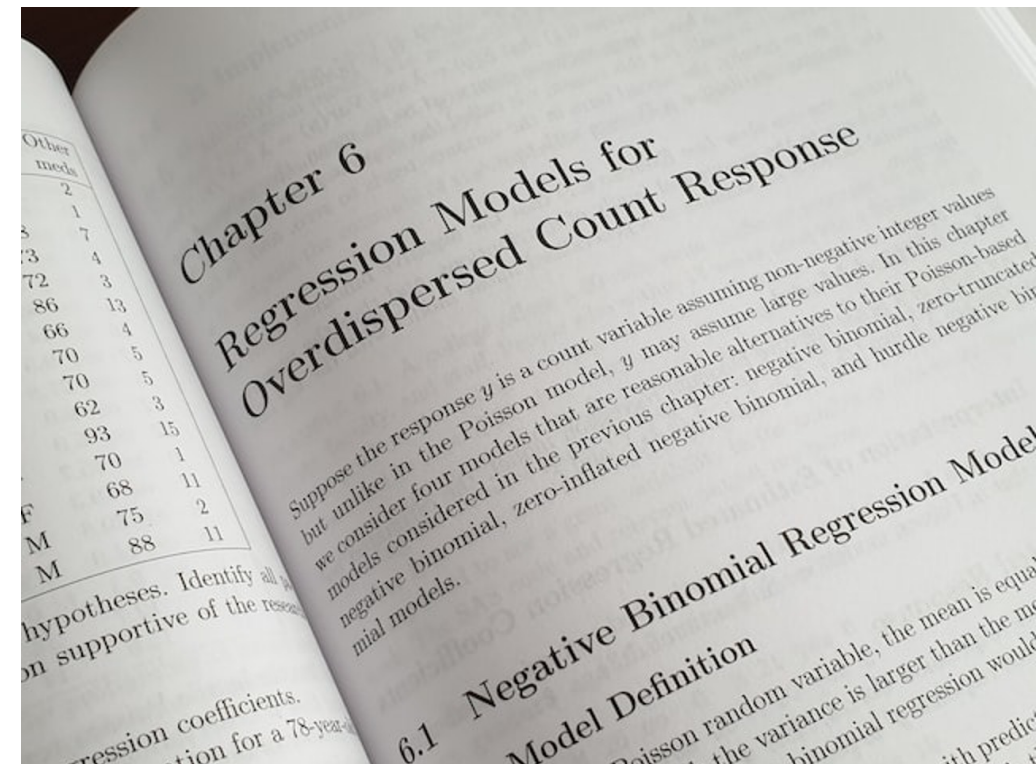


Improve Models with Hyperparameters

- Considerations
 - Pre-processing data
 - These are changes you make to your data before it is passed to the model
 - Scaling features
 - The most common pre-processing step is to scale features so they fall between zero and one.
- Using categories as features
 - 0 or 1 values represent these features in **one-hot vectors**
 - For example, bicycle, skateboard, and car might respectively be (1,0,0), (0,1,0), and (0,0,1).

Activity: Optimize and Save Models

- Go through the Jupyter notebook on the Optimise and Save Models.
- Click on Image to Open the File



Knowledge Check - 4

- You are using scikit-learn to train a regression model from a dataset of sales data. You want to be able to evaluate the model to ensure it will predict accurately with new data. What should you do?
 - A. Use all of the data to train the model. Then use all of the data to evaluate it.
 - B. Train the model using only the features columns, and evaluate it using only the label column.
 - C. Split the data randomly into two subsets. Use one subset to train the model and the other to evaluate it.

Knowledge Check - 5

- You have created a model object using the scikit-learn Linear Regression Class. What should you do to train the model?

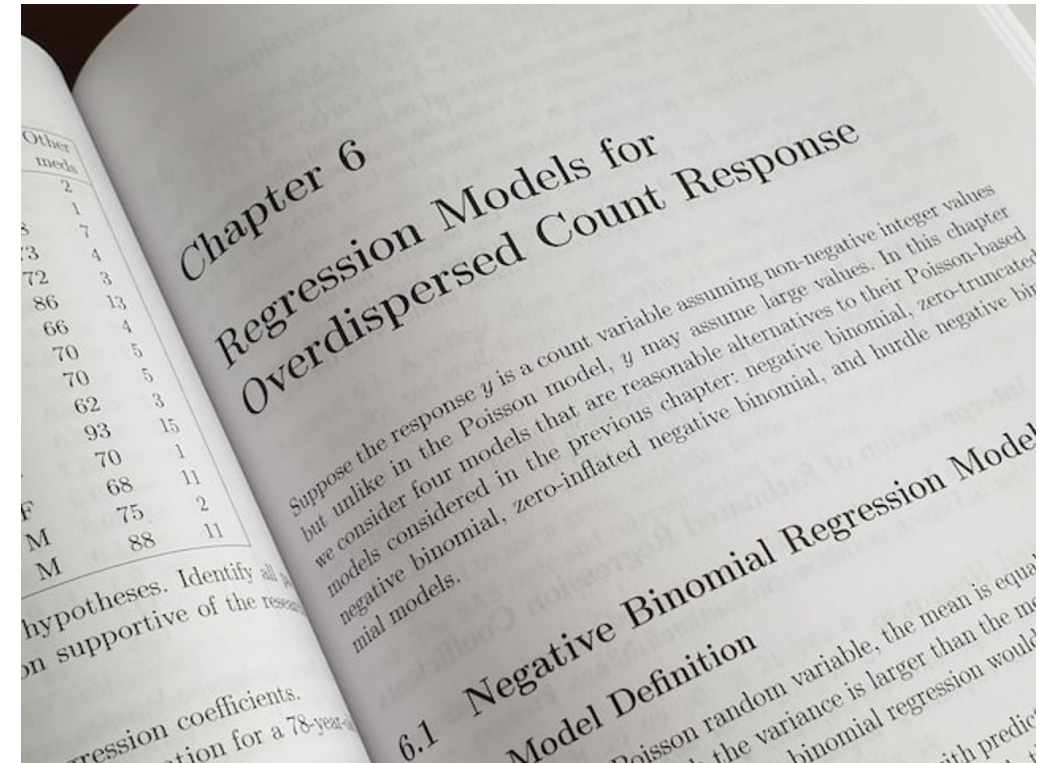
- A. Call the predict() method of the model object, specifying the training feature and label arrays.
- B. Call the fit() method of the model object, specifying the training feature and label arrays.
- C. Call the score() method of the model object, specifying the training feature and test feature arrays

Knowledge Check - 6

- You train a regression model using scikit-learn. When you evaluate it with test data, you determine that the model achieves an R-squared metric of 0.95. What does this metric tell you about the model?
 - A. The model explains most of the variance between predicted and actual values.
 - B. The model is 95% accurate.
 - C. On average, predictions are 0.95 higher than actual values.

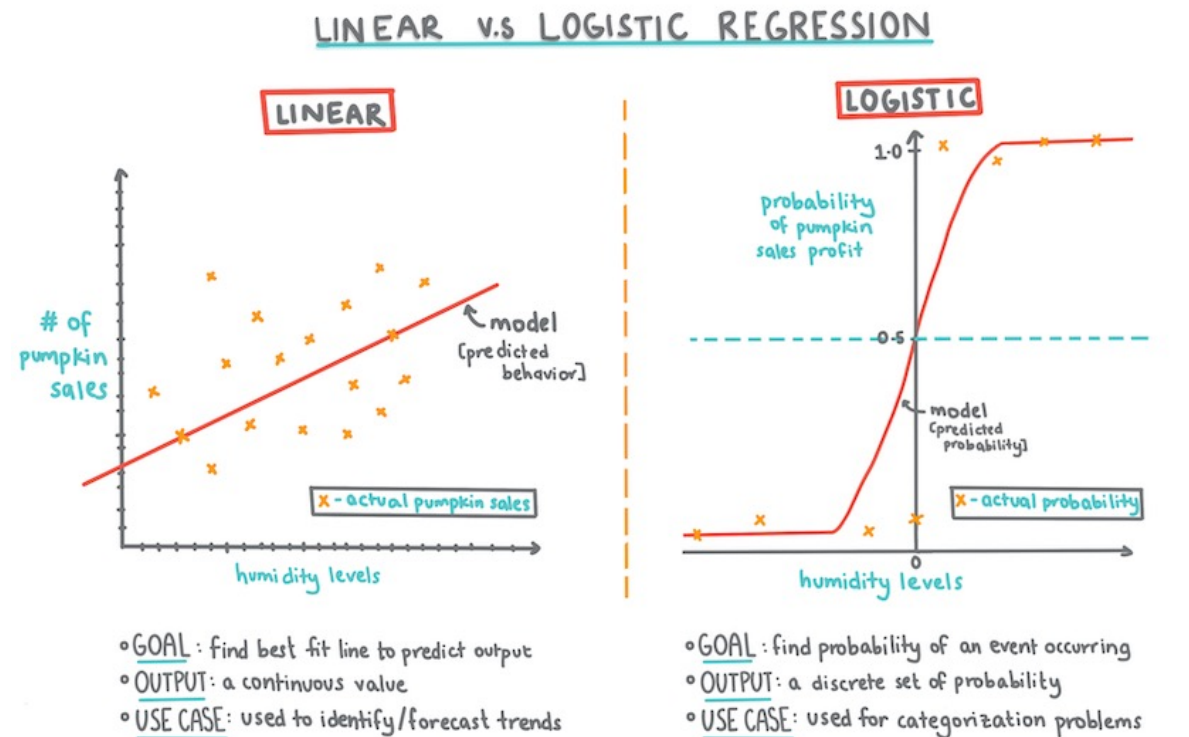
Challenge: Real Estate Regression Challenge

- Go through the Jupyter notebook on the Real Estate Regression Challenge.
- Click on the image to open the file and start the challenge.



Logistic Regression

- logistic regression is a classification method, albeit a linear-based one.
- Logistic regression does not offer the same features as linear regression.
- Logistic regression offers prediction about a binary category whereas linear regression predicts continual values.



@DASANI_DECODED

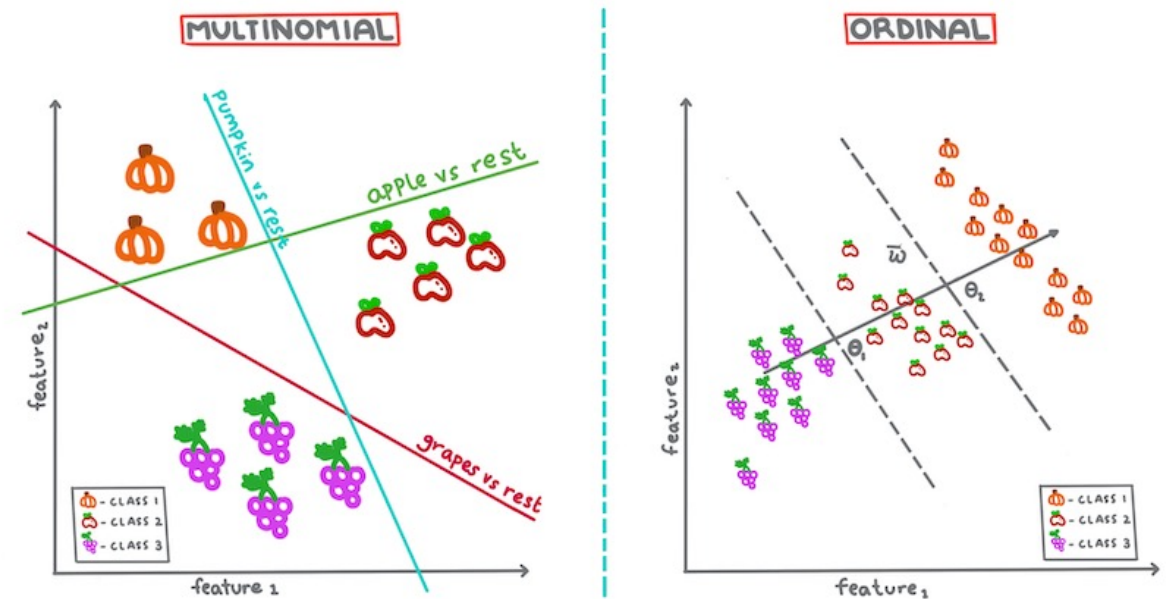
Logistic Regression

- Logistic Regression classification include:
 - Binary
 - Multinomial
 - Ordinal
- Even though this type of Regression is all about 'category predictions',
 - It still works best when there is a clear linear relationship between the dependent variable and the other independent variables
- It's good to get an idea of whether there is any linearity in dividing these variables or not.

Logistic Regression

- Variables DO NOT have to correlate
 - The variables or features do not have to align
- You need a lot of clean data
 - Logistic regression will give more accurate results if you use more data.

MULTINOMIAL v.s ORDINAL LOGISTIC REGRESSION

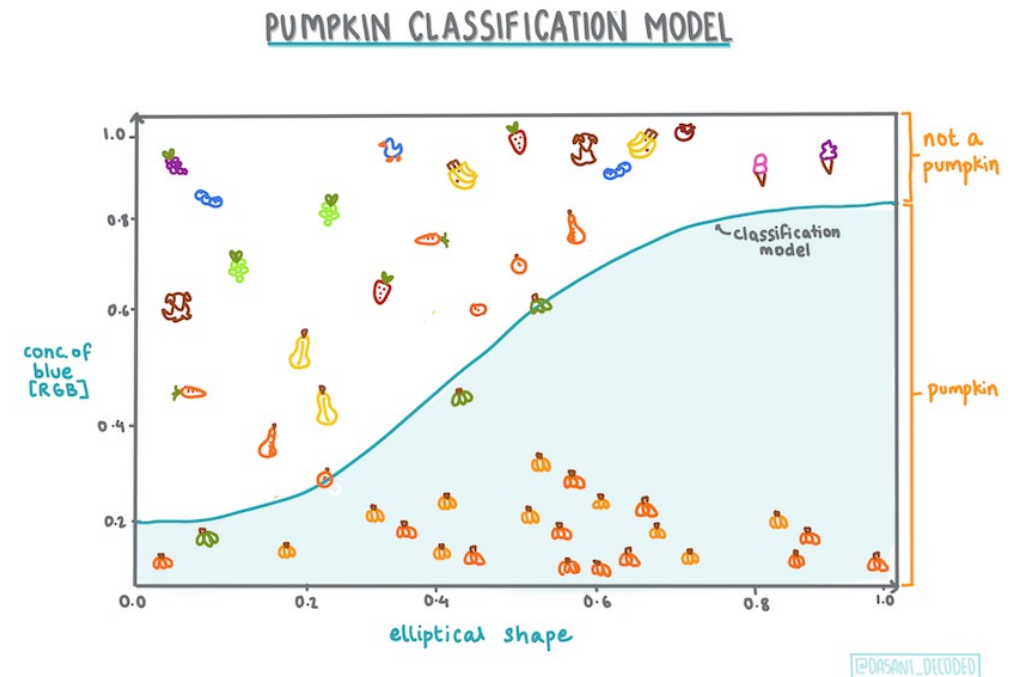


@DASANI_DECODED

Think about the types of data that would lend themselves well to logistic regression

Activity: Logistic Regression

- Go through the Jupyter notebook on Logistic Regression.
- Click on the image to open the file



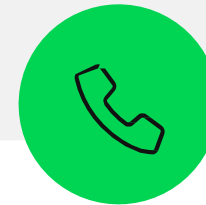
THANK YOU FOR YOUR TIME



RAIL@KNUST.EDU.GH



FECE, COLLEGE OF ENG.,
KNUST, KUMASI GHANA



+233(0)3224-93435

