

# Week 6

FIT5202 Big Data Processing

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Classification Models



# Week 6 Agenda

- Week 5 Review
  - Classification Algorithms
    - Decision Trees
    - Random Forest
    - Logistic Regression
  - Model Evaluation
    - Confusion Matrix
    - ROC Curve
  - Tutorial Use Case
    - Bank : Will customers subscribe?
- Assignment Project Exam Help
- <https://powcoder.com>
- Add WeChat powcoder

# Random Forest

- Use **ensemble** approach
  - The outcome of the model
    - **Majority voting (mode)** (for **classification**)
    - **Mean** of all outcomes (for **regression**)
- Generalise the model
  - Build multiple **different (uncorrelated) trees**
  - **Avoid overfitting** issue found in decision tree
- Use generalisation technique
  - **Bagging (bootstrapping)** – Randomise (with replacement) a different dataset (from the training dataset) used for training each tree.
  - Each tree uses **a random subset of features** for splitting nodes.

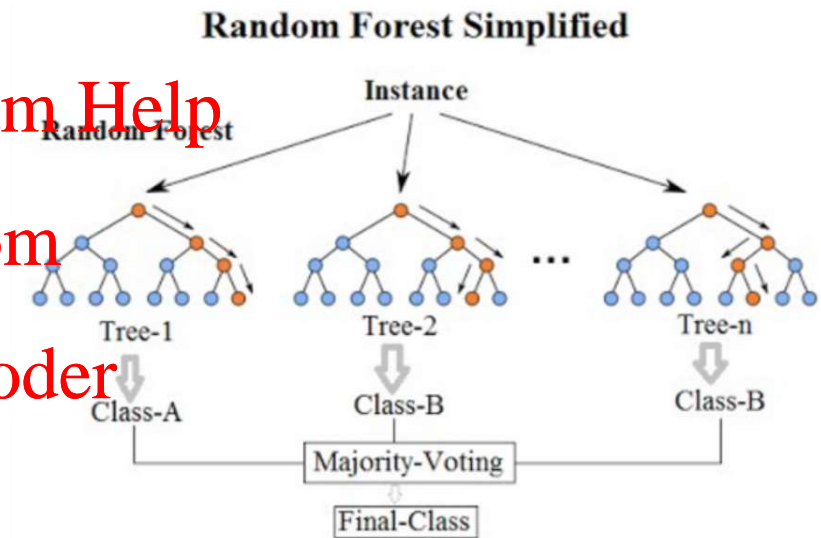


image: <https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d>

# Logistic Regression

- Based on **linear model approach**
  - Instead of predicting continuous target variable,
  - Logistic regression predicts **categorical target variable** (e.g. binary classification)
- Define the **hyperplane** (decision boundary) used to classify data (e.g. to separate the two classes in the data in case of binary classification)

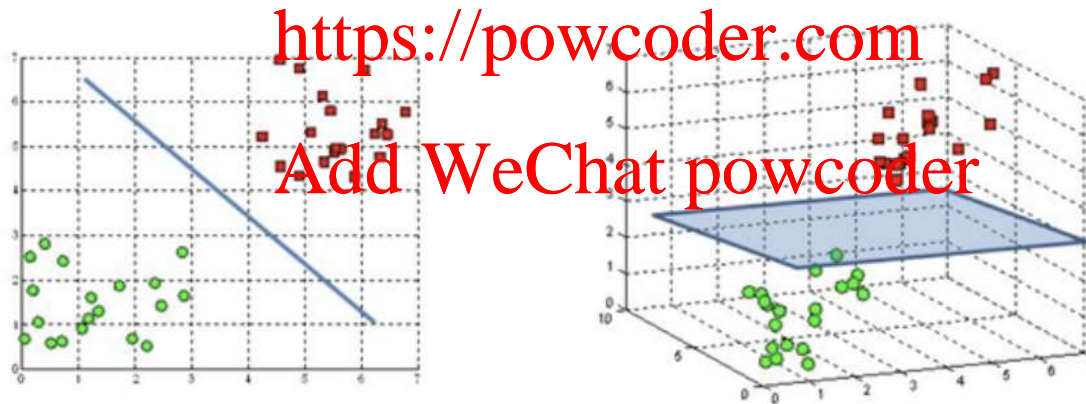
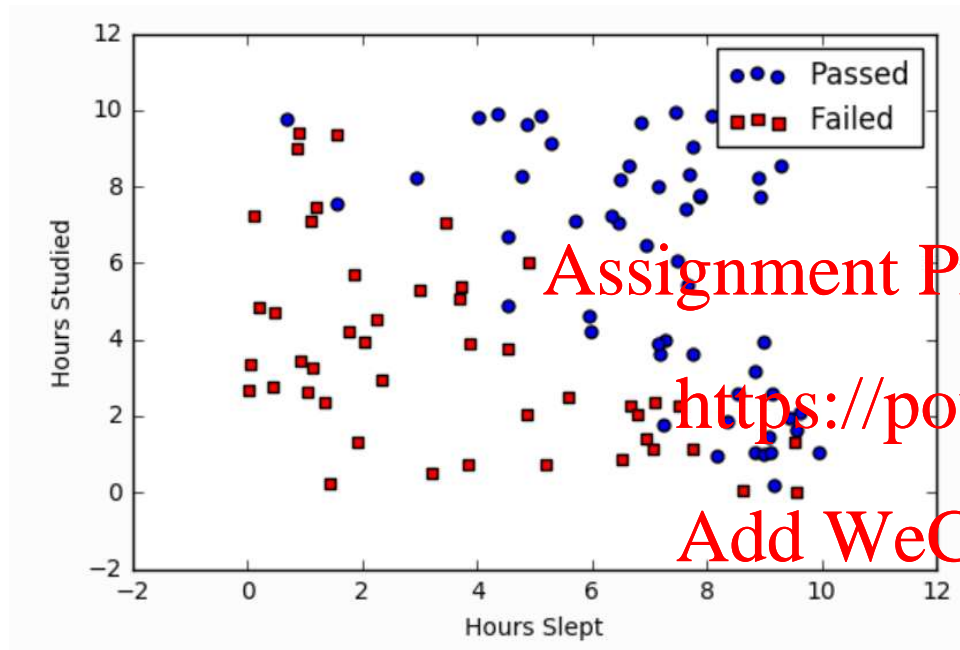


image: [https://www.quora.com/What-is-a-hyperplane-in-machine-learning?top\\_ans=198420733](https://www.quora.com/What-is-a-hyperplane-in-machine-learning?top_ans=198420733)

# Example: Logistic Regression



Assignment Project Exam Help

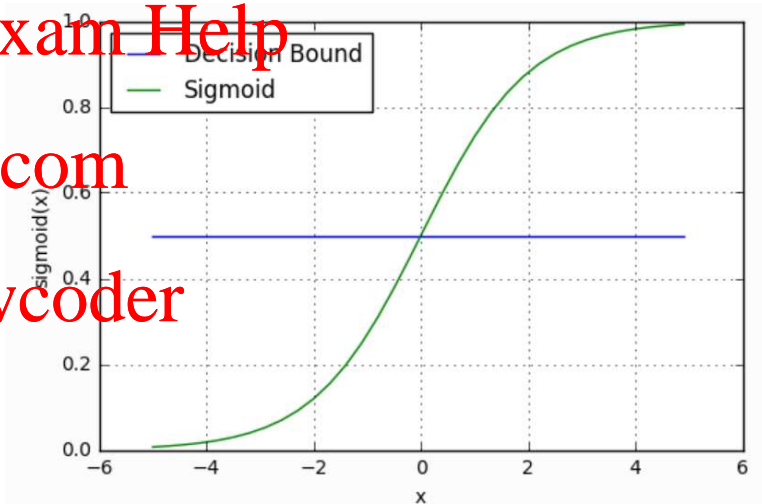
<https://powcoder.com>

Add WeChat powcoder

- **2 features:** Hours slept, Hours studied
- **2 classes:** Passed and Failed

$$z = W_0 + W_1 \text{Studied} + W_2 \text{Slept}$$

$$P(\text{class} = 1) = \frac{1}{1 + e^{-z}}$$



[https://ml-cheatsheet.readthedocs.io/en/latest/logistic\\_regression.html](https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html)

# Evaluating Classifiers

- Threshold Metrics

- Confusion Matrix

- True Positive, True Negative, False Positive, False Negative
    - Accuracy, Precision, Recall (sensitivity) and F1-score

- Ranking Metrics

- ROC Curve

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Classification **accuracy** is almost universally **inappropriate** for imbalanced classification.

# Choosing the right model?

- Understand characteristics of your data?
- Understand characteristics of the model?
- Meets business goals?
- How accurate is the model?
- How explainable is the model?
- How fast is the model?
- How scalable is the model?

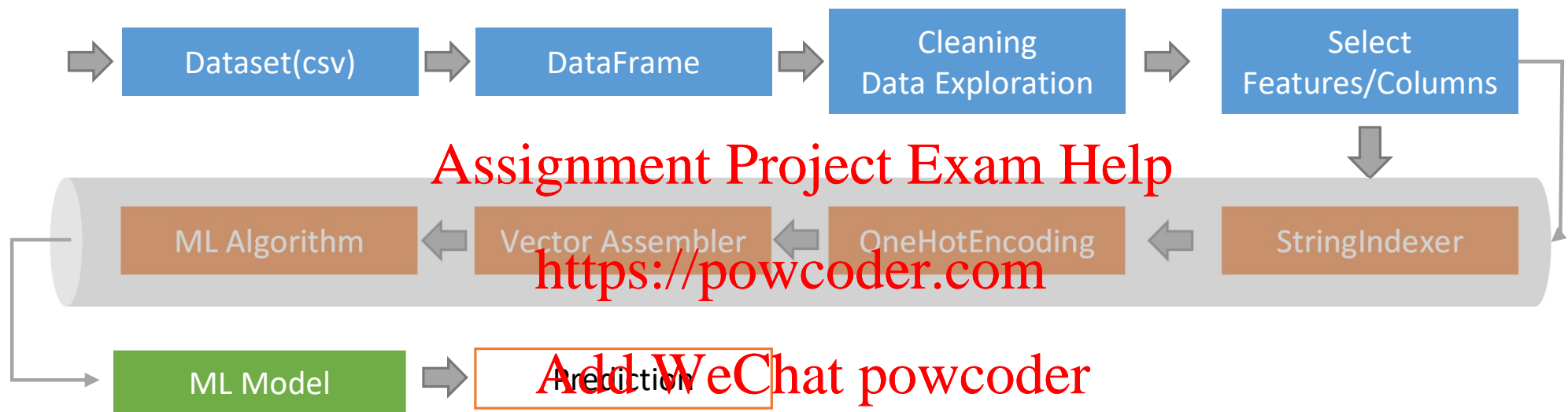
Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

<https://hackernoon.com/choosing-the-right-machine-learning-algorithm-68126944ce1f>

# Bank Use Case: Will the customers subscribe?





**Thank You!**

See you next week.

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder