2025

4Geeks Academy: data science cohort 12

DAY 15: YOUR FIRST ML ALGORITHM

TODO

MACHINE LEARNING

Basics of training machine learning models, what is logistic regression

ALGORITHM OPTIMIZATION PROJECT

Submit Algorithm Optimization Project in for Machine Learning (Algorithm optimization module), if you haven’t done so already

LOGISTIC REGRESSION PROJECT

Work on Logistic Regression Project Tutorial (Your first ML Algorithm module), plan to finish MVP before class Wednesday

TOPICS

01 MACHINE LEARNING

02 TRAINING ML MODELS

03 LOGISTIC REGRESSION

MACHINE LEARNING

WHAT

- Set of techniques and statistical algorithms
- Can ‘learn’ from data
- Goal is to generalize to unseen data, i.e. make predictions

WHY

- Is automatable, robust to different datasets
- Does not require a priori knowledge of relationship between input and output
- Powerfull: can identify higher order relationships in large datasets

HOW

Scikit-learn: open source Python machine learning library, initial release 2007, currently over 32 thousand commits on GitHub

- GitHub repository: scikit-learn
- Official documentation: scikit-learn.org/stable
- PyPI package: scikit-learn

TRAINING ML MODELS

DATA PREPARATION

- Clean: remove redundant & irrelevant data, handle missing data
- Encode: convert strings or objects to numbers
- Improve: scale, normalize etc

FEATURE ENGINEERING

- Choose best features (or use all of them)
- Transform existing features
- Make new features

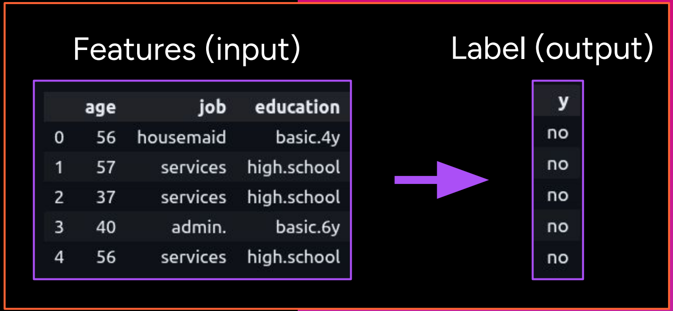
MODEL SELECTION

- Try different model types
- Hyperparameter optimization: tune the model
- Go back and try different data prep/feature engineering

MODEL EVALUATION

Score the model on held-out test data to see how well it has learned to make predictions on new data

LOGISTIC REGRESSION



Features (input)

	age	job	education
0	56	housemaid	basic.4y
1	57	services	high.school
2	37	services	high.school
3	40	admin.	basic.6y
4	56	services	high.school

Label (output)

y
no
no
no
no
no

WHAT

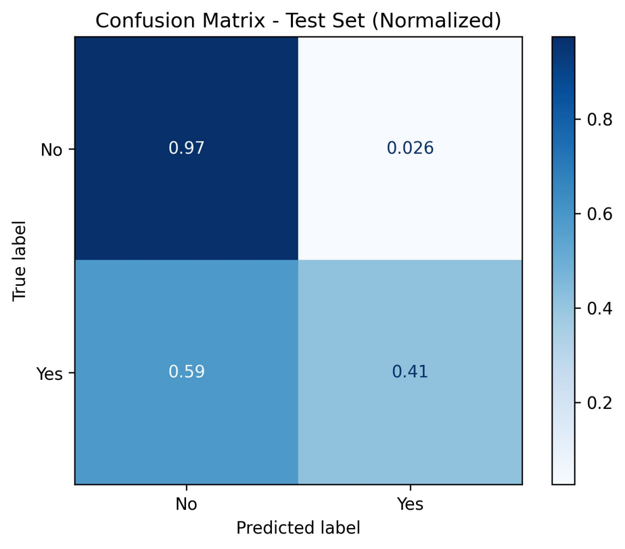
Classification model: outputs the probability that each data point belongs to each of two or more groups

HOW

- Encode string variables to number with `OrdinalEncoder()`
- Split data into training and testing datasets with `train_test_split()`
- Train `LogisticRegression()` model
- Tune hyperparameters with `GridSearchCV()`

EVALUATION

- Evaluate model on test set (data it has not been trained on)
- Overall accuracy percentage is often not a good metric for classification (why?)
- Confusion matrix best way to 'see' how the model is doing



Confusion Matrix - Test Set (Normalized)