

K-Nearest Neighbors Regressor

(CarPrice.csv)

source code : 6610402230_knn.ipynb

1. การทดลองสร้าง machine learning model รูปแบบ regression จาก algorithm k-nearest neighbors เพื่อทำนายราคารถยนต์ด้วยชุดข้อมูล CarPrice.csv มีขั้นตอนดังนี้
2. เตรียมพร้อม environment โดย import package ที่จำเป็นและตั้ง random seed เพื่อกำหนดตัวแปรควบคุม ให้การทดลองแต่ละครั้งสามารถวัดค่าได้อย่างถูกต้อง และมีค่าเหมือนเดิมจากตัวแปรสุ่ม
3. นำเข้าข้อมูล CarPrice.csv เป็น DataFrame
4. Pre-processing data ด้วย
 - drop nan values
 - encode categorical feature ด้วย label-encoding
 - scale ค่าของ feature เพื่อใช้หา k-NN distance ด้วย standard scaler
 - แบ่งข้อมูลชุด train 80% และ test 20 %
5. นำข้อมูลมาฝึกฝน model โดยมี factors ที่ควบคุมดังนี้
 - random_state=seed_value (fixed ค่าตัวแปรสุ่ม)
 - algorithm ในการหา distance ของ neighbors คือ kd-tree
 - วัดระยะทาง distance ด้วย euclidean distance
 - ทดลองหาค่า k-Nearest Neighbors ที่ดีที่สุดในช่วงตั้งแต่ 1 ถึง 100 ด้วยการทำ cross_val_score กับชุดข้อมูล test set โดยให้ weight กับ neighbors แต่ละตัวแบบเท่า “uniform” และวัดประสิทธิภาพด้วย mean absolute error แล้วนำค่า k ที่มี error น้อยที่สุดที่ได้ไปใช้ฝึกฝน
6. วัดประสิทธิภาพโดยมีผลการทดลองดังนี้
 - การฝึกฝนโมเดล ที่ใช้ค่า k จากการทดลองและตัวแปรควบคุมข้างต้นและเลือกใช้ weight แบบ distance เพื่อให้ความสำคัญกับ Neighbors ที่มีความใกล้เคียงชุดข้อมูลมากกว่าทำให้ได้ Accuracy Score มากขึ้น โดยการวัดประสิทธิภาพใช้ model.score(X_test,y_test) ใช้ metric algorithm R^2 ที่หากข้อมูลมีค่า predict ออกจาก mean มากๆ จะทำให้ได้ค่าติดลบ และดูว่าโมเดลมีการทำนายที่ผิดพลาดไปมากเพียงใดจาก metric algorithm แบบ mean absolute error โดยพบว่า
 - R^2 Accuracy Score : 0.59
 - Mean Absolute Error : 171143.28438