Lab1: Experiment Report

6440126922 นิปุณ อังควิชัย

จากการสร้างแบบจำลอง Multiple Linear Regression ทำนายค่าใช้จ่ายทางการแพทย์ที่เรียกเก็บโดยประกันสุขภาพ (Amount) โดยดูจากข้อมูล Age Sex BMI Children Smoker Region ของลูกค้าแต่ละคน หลังการผ่าน Data preprocessing ด้วยการเช็ก missing value และใช้ one-hot encoding แปลงค่าตัวแปรที่เป็น categorical data (Sex, Smoker, Region) ผลลัพธ์ของแบบจำลองจาก Evaluation metrics มีดังนี้

ผลลัพธ์ของแบบจำลองจาก Evaluation metrics

Metrics	โมเดลแรกที่ไม่ได้แปลงค่าตัวแปรตาม* lab1-dm-v1.ipynb	โมเดลที่สองที่ทำการแปลงค่าตัวแปรตามด้วยการ take log lab1-dm-v2.ipynb
R-Square	0.748	0.818
MAE	3.62	0.116
MSE	7.54	0.029
RMSE	3.77	0.170

หมายเหตุ: เมื่อได้ผลลัพธ์จากโมเดลแรกมาแล้ว นำค่า error มา take log เพื่อให้ผลลัพธ์อยู่ในหน่วยเดียวกันและสามารถเปรียบเทียบ performance กับโมเดลที่สองที่มีการแปลงค่าตัวแปรตามตอนเทรนได้

วิเคราะห์และสรุปผลโมเดล

R-Square เป็นการคำนวณสัดส่วนของความผันแปรในตัวแปรตามที่สามารถทำนายได้จากตัวแปรต้น โดยจะมีช่วงอยู่ที่ 0 ถึง 1 ยิ่งเข้าใกล้หนึ่งคือการที่โมเดลสามารถอธิบายความผันแปรของตัวแปรตามได้เป็นอย่างดีจากตัวแปรต้นที่มี ทำให้สามารถวิเคราะห์ได้ว่าโมเดลของเราสามารถอธิบายความผันแปรของค่าตัวแปรตามได้แค่ไหน จากผลลัพธ์ที่เราได้จะเห็นว่าโมเดลที่สองที่มีการแปลงค่าตัวแปรตามด้วยการ take log เป็นโมเดลที่สามารถอธิบายความผันแปรของตัวแปรตามได้มากกว่า จากการที่โมเดลที่สองมีค่า

R-Square สูงกว่าโมเดลแรก แต่ค่า R-Square ไม่ได้แสดงถึงประสิทธิภาพของโมเดลในการทำนายข้อมูลใหม่ เราจึงต้องตรวจสอบ Evaluation metrics อื่นๆ ที่สามารถวัดประสิทธิภาพของโมเดลในการทำนายข้อมูลได้อย่าง MAE MSE และ RMSE ซึ่ง 3 metrics นี้เป็นการคำนวณค่าความคลาดเคลื่อน (error) ยิ่งใกล้ศูนย์มากเท่าไหร่ก็แสดงว่าโมเดลมีข้อผิดพลาดที่น้อย

MAE เป็นการคำนวณค่าเฉลี่ยของค่าสัมบูรณ์ความแตกต่างระหว่างข้อมูลที่ทำนายและข้อมูลที่สังเกต เป็นการคำนวณที่ให้ความสำคัญกับความคลาดเคลื่อนทุกแบบอย่างเท่ากัน และไม่อ่อนไหวต่อค่านอกเกณฑ์ จากผลลัพธ์ที่เราได้จะเห็นได้ว่าโมเดลที่สองมีประสิทธิภาพมากกว่าโมเดลแรกในการวัดผลโมเดลแบบ MAE เกือบ 30 เท่า มีค่า MAE 0.116 มีค่าอยู่ในตำแหน่งทศนิยมตำแหน่งที่ 1 ซึ่งเข้าใกล้ศูนย์มาก

MSE เป็นการคำนวณค่าเฉลี่ยของค่ายกกำลังสองของความแตกต่างระหว่างข้อมูลที่ทำนายและข้อมูลที่สังเกต เป็นการคำนวณที่ขยายความคลาดเคลื่อนที่มีขนาดใหญ่ เนื่องจากมีการยกกำลังสองค่าความแตกต่าง ทำให้มีความอ่อนไหวต่อค่านอกเกณฑ์ จากผลลัพธ์ที่เราได้จะเห็นได้ว่าโมเดลที่สองมีประสิทธิภาพมากกว่าโมเดลแรกในการวัดผลโมเดลแบบ MSE เกือบ 200 เท่า ผลลัพธ์มีความแตกต่างกันมาก อาจเกิดจากการที่ข้อมูลของเรามีค่านอกเกณฑ์มาก เนื่องจากการวัดประสิทธิภาพโมเดลแบบ MSE อ่อนไหวต่อค่านอกเกณฑ์เนื่องจากมีการยกกำลังสองเพื่อขยายน้ำหนักให้ความคลาดเคลื่อนที่เกิดขึ้น

RMSE เป็นการใส่รูทค่า MSE ทำให้ได้ค่าที่กลับมาอยู่ในหน่วยเดียวกันกับตัวแปรตาม ทำให้สามารถตีความได้ง่ายกว่า MSE จากผลลัพธ์ที่เราได้จะเห็นได้ว่าโมเดลที่สองมีประสิทธิภาพมากกว่าโมเดลแรกในการวัดผลโมเดลแบบ MSE เกือบ 30 เท่า ซึ่งมีความใกล้เคียงกับการวัดผลโมเดลแบบ MAE

ในภาพรวมถือว่าโมเดลที่สองมีประสิทธิภาพในการทำนายข้อมูลที่ดีกว่าโมเดลแรกในการวัดผลโมเดลทุกรูปแบบ (Goodness of fit) และดีกว่าในการวัดความเหมาะสมของโมเดล (Goodness of fit)