**Experiment and create the best regression model for predicting daily**

**Answer**

Table

Description automatically generated

**รูปที่ 1 Correlation ของข้อมูลที่สัมพันธ์กับ Target**

เริ่มแรกทำการดู Correlation ของข้อมูล เพื่อดูว่ามีตัวแปรไหนบางที่สัมพันธ์กันในข้อมูล Target ของเรา เพื่อใช้สำหรับในการเลือก Feature สำหรับมาทำ Model จากรูปที่ 1

Table

Description automatically generated

**รูปที่ 2 Correlation ของข้อมูลที่สัมพันธ์กับ Target ที่ทำการเลือก**

หลักจากนั้นทำการกรองเอาค่า Correlations ที่มากกว่า 0.6 และ น้อยกว่า -0.6 จะได้ดังรูปที่2 เพื่อใช้เป็นตัวเลือกสำหรับในการสร้าง Model

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

**รูปที่ 3 แสดง Scatter plot ของ Non-urgent-order กับ target (ซ้าย) ,  
Urgent-order กับ target (ตรงกลาง) และ Order-type-A กับ target (ขวา)**

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

**รูปที่ 4 แสดง Scatter plot ของ Order-type-B กับ target (ซ้าย) ,  
Order-type-C กับ target (ตรงกลาง) และ Banking-order-1 กับ target (ขวา)**

โดยจากการทดลอง จะเห็นได้ว่าบางความสัมพันธ์จะมี Outlier อยู่มากเลยทำการ Outlier ออกด้วย IQR จากการทดลองได้ทำการเอา Outlier Urgent-order ออกอยู่ค่าเดียวเพราะส่งผลต่อ Correlations ในทางที่ดีขึ้น จากรูปที่ 3

Table

Description automatically generated

**รูปที่ 5 Correlation ของข้อมูลที่สัมพันธ์กับ Target ที่ทำการเลือก หลังจากเอา Outlier ออก**

โดยเมื่อนำค่า Outlier ออกทำให้ Correlation บางค่ามีค่าลดลงแต่ ค่าส่วนมากจะเพิ่มขึ้น ดังนั้นเลยได้ใช้ข้อมูลนี้ต่อไปในการ Linear Regression model ต่อไป จากรูปที่ 5

Text

Description automatically generated

**รูปที่ 6 model ที่ 1**

โดยสมการของ model ที่ 1 จากรูปที่ 6 จะได้เป็น target = 0.644nonUrgentOrder + 0.7313orderTypeB + 0.0001242bankingOrder2 + 0.9824 ซึ่งจะได้ค่า S หรือ Sum error square เท่ากับ 27711.84

Text

Description automatically generated

**รูปที่ 7 model ที่ 2**

โดยสมการของ model ที่ 2 จากรูปที่ 7 จะได้เป็น target = 0.648725nonUrgentOrder + 0.75321orderTypeB + 98.18791 ซึ่งจะได้ค่า S หรือ Sum error square เท่ากับ 28217.819 จะสามารถสังเกตุได้ว่า ค่า Adjusted R-squared ไม่ได้เพิ่มขึ้นและ S ก็เยอะมากขึ้นด้วย ซึ่งอาจจะหมายความว่า อาจจะไม่ความจำเป็นที่ต้องบวกสมการ

Text

Description automatically generated

**รูปที่ 8 model ที่ 3**

โดยสมการของ model ที่ 3 จากรูปที่ 8 จะได้เป็น target = 1.15474nonUrgentOrder + 100.15420 ซึ่งจะได้ค่า S หรือ Sum error square เท่ากับ 55157.594 โดยสามารถสังเกตได้ว่า 1 ตัวแปรไม่เพียงพอต่อการสร้าง model ถึงแม้ ตัวแปรนี้จะสัมพันธ์กับ Target ของเรามากที่สุด

Text

Description automatically generated

**รูปที่ 9 model ที่ 4**

โดยสมการของ model ที่ 4 จากรูปที่ 9 จะได้เป็น target = 8.972e-07nonUrgentOrder + 1.393e-04bankingOrder2 + 8.972e-07bankingOrder2\*nonUrgentOrder + 1.134e+02 ซึ่งจะได้ค่า S หรือ Sum error square เท่ากับ 51745.832

**Text

Description automatically generated**

**รูปที่ 10 model ที่ 5**

โดยสมการของ model ที่ 5 จากรูปที่ 10 จะได้เป็น target = 0.83156nonUrgentOrder + 0.9105554orderTypeB - 0.0009499nonUrgentOrder\*orderTypeB + 76.2444057

ซึ่งจะได้ค่า S หรือ Sum error square เท่ากับ 27262.60 ซึ่งจะสามารถสังเกตุเทียบจาก model ที่ 4 ได้ เมื่อทำการเปลี่ยนตัวแปรที่มี ค่า Correlations สูงขึ้นแล้วนำมา Interaction กัน เราจะได้ ค่า Adjusted R-squared ที่สูงกว่าด้วย ดีกว่าการเอาไป additional กัน

Text

Description automatically generated with medium confidence

**รูปที่11 model ที่ 6**

โดยสมการของ model ที่ 6 จากรูปที่ 11 จะได้เป็น target = 0.7836656

nonUrgentOrder + 0.8789629orderTypeB + 0.0001012bankingOrder2 -0.0008668nonUrgentOrder\*orderTypeB + 78.2054969 ซึ่งจะได้ค่า S หรือ Sum error square เท่ากับ 26933.91 ซึ่งเมื่อทดลองนำ ค่าที่ Correlation สูง ๆ มาบวกเพิ่ม ก็ไม่ได้ให้ผลที่ต่างกันมากจากเดิม เมื่อเทียบกับ model ที่ 5

**Text

Description automatically generated**

**รูปที่ 10 model ที่ 7**

โดยสมการของ model ที่ 7 จากรูปที่ 10 จะได้เป็น target = 0.7519976 nonUrgentOrder + 0.9657336orderTypeB + 0.5080484orderTypeA - 0.0011624orderTypeB\*nonUrgentOrder + 61.490982 ซึ่งจะได้ค่า S หรือ Sum error square เท่ากับ 23923.05 ซึ่งจะสามารถสังเกตุเทียบจาก model ที่ 6 ได้ เมื่อทำการเปลี่ยนตัวแปรที่มี ค่า Correlations ต่ำกว่า แต่ว่ากลับจะได้ S ที่มากขึ้น นั้นแปลว่าตัวแปรนี้อาจเป็นตัวแปรที่มีความสำคัญแต่ไม่สามารถแสดงในรูปของความสัมพันธ์ได้ ถ้านำมา interaction อาจจะให้ผลที่ดีขึ้น

**Table

Description automatically generated with medium confidence**

**รูปที่11 model ที่ 8**

โดยสมการของ model ที่ 8 จากรูปที่ 11 จะได้เป็น target = - 1.296e-16

nonUrgentOrder + orderTypeB + orderTypeA + orderTypeC - 1.326e-18

nonUrgentOrder\*orderTypeB + nonUrgentOrder\*orderTypeC ซึ่งจะได้ค่า S หรือ Sum error square เท่ากับ 1.57e-25 โดยเมื่อลองนำ ตัวแปลที่ Correlation ไม่สูงมากมา Interaction กัน เราจะได้ model ที่มีค่า Adjusted R-squared เท่ากับ 1 ซึ่งเป็น model ที่ fit กับข้อมูลมาก ๆ จน Overfit โดยสามารถสังเกตุได้อีกย่างคือ ตัว Feature orderTypeB orderTypeA orderTypeC มีค่าความชันเป็น 1 ซึ่งมันอาจหมายความว่าตัวแปรเหล่านี้เป็นตัวแปรที่กำหนดสมการหลัก ในแปรผันตาม Target ของเรา

Text, table

Description automatically generated with medium confidence

**รูปที่12 ทำ Train test split เพื่อทดสอบ model ที่ 8**

โดยเมื่อลองทำ Train test split ก็ได้ผลลัพธ์ที่ดีมาก จากการทดลอง รูปที่12 ซึ่งจากการทดลองนี้ข้อมูลอาจจะน้อยเกินไป จนอาจจะไม่ reflect ถึง data ที่แท้จริงได้ ทำให้ model มีค่าที่ Fit ขนาดนี้

**Why is sometime adding predictors do not help prediction?**

**Answer** จากาการทดลองเราจะเห็นได้ว่า บางครั้งการที่เราเพิ่มตัวแปรที่มี Correlations สูงก็อาจจะไม่ได้ทำให้ model มีประสิทธิภาพสูงเสมอไป แต่เราต้องหาตัวแปรที่เป็น insight ของข้อมูลของเรา จะช่วยเพิ่มประสิทธิภาพของ model ได้ดีมากกว่า ที่จะช่วยในการ predict จริง ๆ