Finding the Relationship of Data with the Label Attribute

ข้อที่ 1

บทนำ

ตัวแปรที่นำมาทดสอบสมมติฐาน Test of Two Means : Independent Sample คือ เพศ (Gender) กับ จำนวนธุรกรรมที่จ่าย (Total_Trans_Ct)

เหตุผลที่เลือก 2 ตัวแปรนี้เกิดจากการตั้งข้อสังเกตหรือปัญหาที่ว่า ค่าเฉลี่ยประชากรของจำนวนธุรกรรมที่จ่ายระหว่างเพศหญิงและเพศชาย มีค่าเท่ากันหรือไม่ และทั้ง 2 ตัวแปรนั้นเป็นอิสระต่อกันทำให้เหมาะสมต่อการนำมาทดสอบสมมติฐานแบบดังที่กล่าวไปข้างต้น

ผลการวิเคราะห์

1. ตั้งสมมติฐานหลักและสมมติฐานรองดังนี้

   Ho : ค่าเฉลี่ยประชากรของจำนวนธุรกรรมที่จ่ายเพศหญิง = ค่าเฉลี่ยประชากรของจำนวนธุรกรรมที่จ่ายเพศชาย

   Ha : ค่าเฉลี่ยประชากรของจำนวนธุรกรรมที่จ่ายเพศหญิง ≠ ค่าเฉลี่ยประชากรของจำนวนธุรกรรมที่จ่ายเพศชาย

2. คำสั่งที่ใช้ในโปรแกรม R Studio

   มีการประกาศตัวแปร เพื่อเก็บข้อมูลในแต่ละคอลัมน์ และใช้ Function t.test() ในการประมวลผล เพื่อดูความแตกต่างของค่าเฉลี่ยประชากร

```r
gender <- Final$Gender
total_trans <- Final$Total_Trans_Ct
t.test(total_trans ~ gender, var.equal=TRUE)
```

รูปภาพที่ 1 : แสดงคำสั่งที่ใช้ในการประมวลผล

3. ผลที่ได้จากโปรแกรม R Studio

จากการประมวลผล ได้ผลลัพธ์ที่แสดงค่า p-value = 0.8291 ที่ระดับนัยสำคัญ 0.05 นอกจากนั้น มีการแสดงค่าเฉลี่ยประชากรของเพศหญิงเท่ากับ 69.79730 และค่าเฉลี่ยประชากรของเพศชายเท่ากับ 69.27451

```
        Two Sample t-test

data:  total_trans by gender
t = 0.21611, df = 197, p-value =
0.8291
alternative hypothesis: true difference in means
between group F and group M is not equal to 0
95 percent confidence interval:
 -4.247911  5.293486
sample estimates:
mean in group F mean in group M
       69.79730        69.27451
```

รูปภาพที่ 2 : แสดงผลลัพธ์ที่ได้จากการประมวลผล

4. ข้อสรุปของการทดสอบสมมติฐาน

จากผลลัพธ์ที่ได้คือค่า p-value ที่มีค่า 0.8291 นั้นมีค่ามากกว่า ระดับนัยสำคัญที่ 0.05 แสดงว่าไม่สามารถปฏิเสธสมมติฐานหลักได้

บทสรุป

จากการตั้งข้อสังเกตที่ว่า ค่าเฉลี่ยประชากรของจำนวนธุรกรรมที่จ่ายเพศหญิงและเพศชาย มีค่าเท่ากันหรือไม่ พบว่า ค่าเฉลี่ยประชากรของจำนวนธุรกรรมที่จ่าย ทั้ง 2 เพศนั้นมีค่าเท่ากันที่ระดับนัยสำคัญ เท่ากับ 0.05

ข้อที่ 2

บทนำ (แบบจำลองที่ 1)

การพยากรณ์สถานะของลูกค้า (Attrition_Flag) ตัวแปรอิสระคือ มูลค่ารวมของทุกรายการที่จ่าย (Total_Trans_Amt) กับจำนวนธุรกรรมที่ใช้จ่าย (Total_Trans_Ct) เหตุผลที่เลือก 2 ตัวแปรนี้ เพราะตั้งขอสังเกตว่าน่าจะมีความสัมพันธ์กับสถานะของลูกค้า และจะทำให้การพยากรณ์ออกมาได้มีความแม่นยำสูง นอกจากนั้นได้มองว่าตัวแปรอื่นมีค่า 0 เยอะเกินไป หรืออาจจะไม่เกี่ยวข้องกับสถานะลูกค้าจากการวิเคราะห์

ผลการวิเคราะห์ (แบบจำลองที่ 1)

1. ตั้งสมมติฐานหลักและสมมติฐานรองดังนี้

   กำหนดให้ β_1 แทนค่า Slope ของตัวแปรจำนวนธุรกรรมที่จ่าย

   Ho : $\beta_1 = 0$

   Ha : $\beta_1 \neq 0$

   และ

   กำหนดให้ β_2 แทนค่า Slope ของตัวแปรมูลค่ารวมของทุกรายการที่จ่าย

   Ho : $\beta_2 = 0$

   Ha : $\beta_2 \neq 0$

2. คำสั่งที่ใช้ในโปรแกรม R Studio

   มีการประกาศตัวแปร เพื่อเก็บข้อมูลในแต่ละคอลัมน์ และแปลงสถานะลูกค้าให้เป็นเลขโดย Attrited Customer เป็นเลข 0 และ Existing Customer เป็นเลข 1 เพื่อเป็นการเตรียมข้อมูลก่อนการประมวลผล และประมวลผลด้วย Function glm() รวมถึงแสดงผลลัพธ์จากการประมวลผล

   นอกจากนั้นยังได้สร้าง Confusion Matrix เพื่อหาความแม่นยำในการพยากรณ์สถานะของลูกค้า จากตัวแปรอิสระทั้ง 2 ตัว โดยกำหนดจุด cutoff ที่ 0.5

```r
attrition_flag <- Final$Attrition_Flag
total_trans_ct <- Final$Total_Trans_Ct
total_trans_amt <- Final$Total_Trans_Amt
attrition_flag <- ifelse(attrition_flag == "Existing Customer",1,0)
result = glm(attrition_flag ~ total_trans_ct + total_trans_amt,
             family=binomial)
summary(result)

pred <- predict(result,type = "response")
table(pred > 0.5, attrition_flag)
```

รูปภาพที่ 3 : แสดงคำสั่งที่ใช้ในการประมวลผล

3. ผลที่ได้จากโปรแกรม R Studio

จากการประมวลผลได้ค่า p-value ของตัวแปรจำนวนธุรกรรมที่จ่าย เท่ากับ 0.251512 และ ตัวแปรมูลค่ารวมของทุกรายการที่จ่าย เท่ากับ 0.014925 ที่ระดับนัยสำคัญ 0.05

และได้ค่าจากการพยากรณ์สถานะของลูกค้า ได้แก่ True Positive เท่ากับ 160 และ False Positive เท่ากับ 38

```
Coefficients:
                  Estimate Std. Error z value
(Intercept)     -21.049769   5.757662  -3.656
total_trans_ct    0.110148   0.096058   1.147
total_trans_amt   0.004949   0.002033   2.434
                Pr(>|z|)
(Intercept)     0.000256 ***
total_trans_ct  0.251512
total_trans_amt 0.014925 *
---
```

รูปภาพที่ 4 : แสดงผลลัพธ์ที่แสดงความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ

```
      attrition_flag
         0   1
FALSE   38   1
TRUE     0 160
```

รูปภาพที่ 5 : แสดงผลลัพธ์ตาราง Confusion Matrix

4. ความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตาม

    จากผลลัพธ์ค่า p-value ของตัวแปรอิสระจำนวนธุรกรรมที่จ่ายนั้นมีค่าเท่ากับ 0.251512 ซึ่งมีมากกว่าระดับนัยสำคัญที่ 0.05 แสดงว่าไม่สามารถปฏิเสธสมมติฐานหลักได้ และสรุปผลได้ว่า ตัวแปรจำนวนธุรกรรมที่จ่ายไม่มีความสัมพันธ์กับ สถานะของลูกค้า ที่ระดับนัยสำคัญ 0.05

    ค่า p-value ของตัวแปรอิสระมูลค่ารวมของทุกรายการที่จ่ายนั้นมีค่าเท่ากับ 0.014925 ซึ่งมีน้อยกว่าระดับนัยสำคัญที่ 0.05 แสดงว่าสามารถปฏิเสธสมมติฐานหลักได้ และสรุปผลได้ว่า ตัวแปรมูลค่ารวมของทุกรายการที่จ่าย มีความสัมพันธ์กับ สถานะของลูกค้า ที่ระดับนัยสำคัญ 0.05

5. คำนวณความแม่นยำของการพยากรณ์

    จากตาราง Confusion Matrix พบว่ามีการพยากรณ์ค่าได้ True Positive กับ False Positive รวมกันได้ 198 จากจำนวนตัวอย่าง 199 แสดงว่า มีค่าความแม่นยำของการพยาการณ์สถานะลูกค้าอยู่ที่ 99.50%

บทนำ (แบบจำลองที่ 2)

การพยากรณ์สถานะของลูกค้า (Attrition_Flag) ตัวแปรอิสระคือ วงเงินรวม (Credit_Limit) กับจำนวนเดือนที่ใช้บริการ (Months_on_book) เหตุผลที่เลือก 2 ตัวแปรนี้ เพราะตั้งขอสังเกตว่าน่าจะมีความสัมพันธ์กับสถานะของลูกค้า และจะทำให้การพยากรณ์ออกมาได้มีความแม่นยำสูง นอกจากนั้นได้มองว่าตัวแปรอื่นมีค่า 0 เยอะเกินไป หรืออาจจะไม่เกี่ยวข้องกับสถานะลูกค้าจากการวิเคราะห์

ผลการวิเคราะห์ (แบบจำลองที่ 2)

1. ตั้งสมมติฐานหลักและสมมติฐานรองดังนี้

   กำหนดให้ $\beta 1$ แทนค่า Slope ของตัวแปรวงเงินรวม

   Ho : $\beta 1 = 0$

   Ha : $\beta 1 \neq 0$

   และ

   กำหนดให้ $\beta 2$ แทนค่า Slope ของตัวแปรจำนวนเดือนที่ใช้บริการ

   Ho : $\beta 2 = 0$

   Ha : $\beta 2 \neq 0$

2. คำสั่งที่ใช้ในโปรแกรม R Studio

   มีการประกาศตัวแปร เพื่อเก็บข้อมูลในแต่ละคอลัมน์ และแปลงสถานะลูกค้าให้เป็นเลขโดย Attrited Customer เป็นเลข 0 และ Existing Customer เป็นเลข 1 เพื่อเป็นการเตรียมข้อมูลก่อนการประมวลผล และประมวลผลด้วย Function glm() รวมถึงแสดงผลลัพธ์จากการประมวลผล

   นอกจากนั้นยังได้สร้าง Confusion Matrix เพื่อหาความแม่นยำในการพยากรณ์สถานะของลูกค้า จากตัวแปรอิสระทั้ง 2 ตัว โดยกำหนดจุด cutoff ที่ 0.5

```
attrition_flag <- Final$Attrition_Flag
months_on_book <- Final$Months_on_book
credit_limit <- Final$Credit_Limit
attrition_flag <- ifelse(attrition_flag == "Existing Customer",1,0)
result = glm(attrition_flag ~ months_on_book + credit_limit,
             family=binomial)
summary(result)

pred <- predict(result,type = "response")
table(pred > 0.5, attrition_flag)
```

รูปภาพที่ 6 : แสดงคำสั่งที่ใช้ในการประมวลผล

3. ผลที่ได้จากโปรแกรม R Studio

จากการประมวลผลได้ค่า p-value ของตัวแปรวงเงินรวม เท่ากับ 0.239 และ ตัวแปรจำนวนเดือนที่ใช้บริการ เท่ากับ 0.455 ที่ระดับนัยสำคัญ 0.05

และได้ค่าจากการพยากรณ์สถานะของลูกค้า ได้แก่ True Positive เท่ากับ 161 และ False Positive เท่ากับ 38

```
Coefficients:
                Estimate Std. Error
(Intercept)    4.615e-01  1.007e+00
months_on_book 2.071e-02  2.771e-02
credit_limit   6.018e-05  5.112e-05
               z value Pr(>|z|)
(Intercept)      0.458    0.647
months_on_book   0.747    0.455
credit_limit     1.177    0.239
```

รูปภาพที่ 7 : แสดงผลลัพธ์ที่แสดงความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ

	attrition_flag	
	0	1
TRUE	38	161

รูปภาพที่ 8 : แสดงผลลัพธ์ตาราง Confusion Matrix

4. ความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตาม

จากผลลัพธ์ค่า p-value ของตัวแปรอิสระวงเงินรวม มีค่าเท่ากับ 0.239 ซึ่งมีมากกว่าระดับนัยสำคัญที่ 0.05 แสดงว่าไม่สามารถปฏิเสธสมมติฐานหลักได้ และสรุปผลได้ว่า ตัวแปรวงเงินรวม ไม่มีความสัมพันธ์กับ สถานะของลูกค้า ที่ระดับนัยสำคัญ 0.05

ค่า p-value ของตัวแปรอิสระจำนวนเดือนที่ใช้บริการ มีค่าเท่ากับ 0.455 ซึ่งมีมากกว่าระดับนัยสำคัญที่ 0.05 แสดงว่าไม่สามารถปฏิเสธสมมติฐานหลักได้ และสรุปผลได้ว่า ตัวแปรจำนวนเดือนที่ใช้บริการ ไม่มีความสัมพันธ์กับ สถานะของลูกค้า ที่ระดับนัยสำคัญ 0.05

5. คำนวณความแม่นยำของการพยากรณ์

จากตาราง Confusion Matrix พบว่ามีการพยากรณ์ค่าได้ True Positive กับ False Positive รวมกันได้ 199 จากจำนวนตัวอย่าง 199 แสดงว่า มีค่าความแม่นยำของการพยาการณ์สถานะลูกค้าอยู่ที่ 100%

บทสรุป

จากการสร้างแบบจำลองพยากรณ์ สถานะของลูกค้า นั้นมีค่าความแม่นยำของทั้ง 2 แบบจำลองนั้นสูง แต่เมื่อแสดงผลความสัมพันธ์ของตัวแปร พบว่า ส่วนใหญ่ตัวแปรอิสระนั้นจะไม่มีความสัมพันธ์กับ สถานะของลูกค้า ซึ่งอาจจะมองได้ว่าอาจจะเกิดจากการที่ จำนวนข้อมูลที่ใช้ในการสร้างแบบจำลองนั้นน้อยเกินไป ทำให้เกิดความผิดพลาดของการแสดงผล เพราะความจริงแล้วการพยากรณ์ที่ดี ควรขึ้นอยู่กับการมีความสัมพันธ์ต่อกันระหว่างตัวแปรอิสระกับตัวแปรตาม

ข้อที่ 3

บทนำ

Dataset ที่เลือกใช้เกี่ยวกับราคาบ้านกับองค์ประกอบของบ้านต่างๆ โดยต้องการหาความสัมพันธ์ระหว่างราคาบ้าน กับองค์ประกอบอื่นๆที่เกี่ยวข้องกับบ้าน ว่ามีความสัมพันธ์ต่อกันหรือไม่ โดยใช้หลักการ Multiple Linear Regression

ตัวแปรตามได้แก่ ราคาบ้าน (Price) ส่วนตัวแปรอิสระได้แก่ พื้นที่ของบ้าน (Area), จำนวนห้องนอน (Bedrooms), จำนวนห้องน้ำ (Bathrooms), จำนวนชั้นของบ้าน (Stories) และจำนวนรถที่จอดได้ (Parking) โดยเหตุผลที่เลือกตัวแปรอิสระดังที่กล่าวมา เพราะตั้งข้อสังเกตว่าตัวแปรดังกล่าวนั้นจะมีความสัมพันธ์ต่อราคาของบ้าน

ผลการวิเคราะห์

1. ตั้งสมมติฐานหลักและสมมติฐานรองดังนี้

   แบบ Significance of Slope

   กำหนดให้ β_1 แทนค่า Slope ของตัวแปรพื้นที่ของบ้าน

   Ho : $\beta_1 = 0$

   Ha : $\beta_1 \neq 0$

   กำหนดให้ β_2 แทนค่า Slope ของตัวแปรจำนวนห้องนอน

   Ho : $\beta_2 = 0$

   Ha : $\beta_2 \neq 0$

   กำหนดให้ β_3 แทนค่า Slope ตัวแปรจำนวนห้องน้ำ

   Ho : $\beta_3 = 0$

   Ha : $\beta_3 \neq 0$

   กำหนดให้ β_4 แทนค่า Slope ตัวแปรจำนวนชั้นของบ้าน

   Ho : $\beta_4 = 0$

   Ha : $\beta_4 \neq 0$

   กำหนดให้ β_5 แทนค่า Slope ตัวแปรจำนวนรถที่จอดได้

   Ho : $\beta_5 = 0$

   Ha : $\beta_5 \neq 0$

   แบบ Overall Significance

   Ho : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

   Ha : มี β อย่างน้อย 1 ตัว $\neq 0$

2. คำสั่งที่ใช้ในโปรแกรม R Studio

มีการประกาศตัวแปร เพื่อเก็บข้อมูลในแต่ละคอลัมน์ และทำการประมวลผลด้วย Function lm() และแสดงผลลัพธ์จากการประมวลผล

```
price <- Housing$price
area <- Housing$area
bedrooms <- Housing$bedrooms
bathrooms <- Housing$bathrooms
stories <- Housing$stories
parking <- Housing$parking
result <- lm(price ~ area + bedrooms + bathrooms + stories + parking)
summary(result)
```

รูปภาพที่ 9 : แสดงคำสั่งที่ใช้ในการประมวลผล

3. ผลที่ได้จากโปรแกรม R Studio

จากการประมวลผลได้ค่า p-value ของตัวแปรอิสระแต่ละตัว ได้แก่ ตัวแปรพื้นที่ของบ้าน น้อยกว่า 2e-16 ตัวแปรจำนวนห้องนอน เท่ากับ 0.0435 ตัวแปรจำนวนห้องน้ำ น้อยกว่า 2e-16 ตัวแปรจำนวนชั้นของบ้าน เท่ากับ 1.07e-14 ตัวแปรจำนวนรถที่จอดได้ เท่ากับ 2.57e-08

นอกจากนั้นยังแสดงค่า Adjusted R-squared มีค่าเท่ากับ 0.5575 ซึ่งยิ่งมีค่ามาก ยิ่งอธิบาย Explained Variation ได้ดี รวมถึงแสดง p-value โดยรวม น้อยกว่า 2.2e-16

	t value	Pr(>|t|)
(Intercept)	-0.591	0.5548
area	12.448	< 2e-16
bedrooms	2.023	0.0435
bathrooms	9.541	< 2e-16
stories	7.953	1.07e-14
parking	5.652	2.57e-08

รูปภาพที่ 10 : แสดงผลลัพธ์ที่แสดงความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ

```
Residual standard error: 1244000 on 539 degrees of freedom
Multiple R-squared:  0.5616,    Adjusted R-squared:  0.5575
F-statistic: 138.1 on 5 and 539 DF,  p-value: < 2.2e-16
```

รูปภาพที่ 11 : แสดงผลลัพธ์อื่นๆจากการประมวลผล

4. ข้อสรุปของการทดสอบสมมติฐาน

แบบ Significance of Slope

จากผลลัพธ์ค่า p-value ของตัวแปรพื้นที่บ้าน น้อยกว่า 2e-16 ซึ่งมีค่าน้อยกว่าระดับนัยสำคัญที่ 0.05 แสดงว่าสามารถปฏิเสธสมมติฐานหลักได้ และสรุปผลได้ว่า ตัวแปรพื้นที่บ้าน มีความสัมพันธ์กับ ราคาบ้าน ที่ระดับนัยสำคัญ 0.05

ค่า p-value ของตัวแปรจำนวนห้องนอน เท่ากับ 0.0435 ซึ่งมีค่าน้อยกว่าระดับนัยสำคัญที่ 0.05 แสดงว่าสามารถปฏิเสธสมมติฐานหลักได้ และสรุปผลได้ว่า ตัวแปรจำนวนห้องนอน มีความสัมพันธ์กับ ราคาบ้าน ที่ระดับนัยสำคัญ 0.05

ค่า p-value ของตัวแปรจำนวนห้องน้ำ น้อยกว่า 2e-16 ซึ่งมีค่าน้อยกว่าระดับนัยสำคัญที่ 0.05 แสดงว่าสามารถปฏิเสธสมมติฐานหลักได้ และสรุปผลได้ว่า ตัวแปรจำนวนห้องน้ำ มีความสัมพันธ์กับ ราคาบ้าน ที่ระดับนัยสำคัญ 0.05

ค่า p-value ของตัวแปรจำนวนชั้นของบ้าน เท่ากับ 1.07e-14 ซึ่งมีค่าน้อยกว่าระดับนัยสำคัญที่ 0.05 แสดงว่าสามารถปฏิเสธสมมติฐานหลักได้ และสรุปผลได้ว่า ตัวแปรจำนวนชั้นของบ้าน มีความสัมพันธ์กับ ราคาบ้าน ที่ระดับนัยสำคัญ 0.05

ค่า p-value ของตัวแปรจำนวนรถที่จอดได้ เท่ากับ 2.57e-08 ซึ่งมีค่าน้อยกว่าระดับนัยสำคัญที่ 0.05 แสดงว่าสามารถปฏิเสธสมมติฐานหลักได้ และสรุปผลได้ว่า ตัวแปรจำนวนรถที่จอดได้ มีความสัมพันธ์กับ ราคาบ้าน ที่ระดับนัยสำคัญ 0.05

แบบ Overall Significance

ค่า p-value แต่ละตัวนั้นบ่งบอกว่า มีอย่างน้อย 1 ตัวแปรที่มีค่าน้อยกว่าระดับนัยสำคัญที่ 0.05 แสดงว่าสามารถปฏิเสธสมมติฐานได้ และสรุปโดยภาพรวมได้ว่าตัวแปรอิสระนั้น มีความสัมพันธ์กับ ราคาบ้าน

บทสรุป

จากการตั้งข้อสังเกตที่ว่าตัวแปรอิสระทั้ง 5 ตัวนั้น มีความสัมพันธ์กับตัวแปรตามที่เป็นราคาบ้านหรือไม่ พบว่า ตัวแปรอิสระทุกตัว มีความสัมพันธ์กับราคาบ้าน จากการอ้างอิงตัวเลขที่ได้จากการประมวลผล