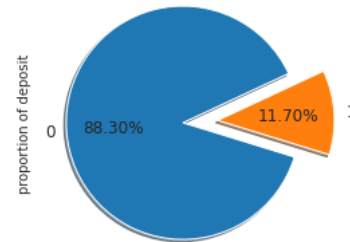
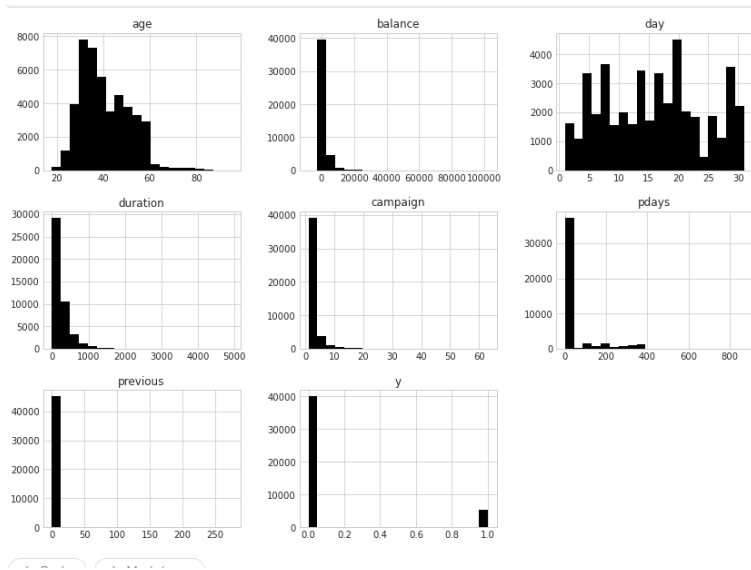


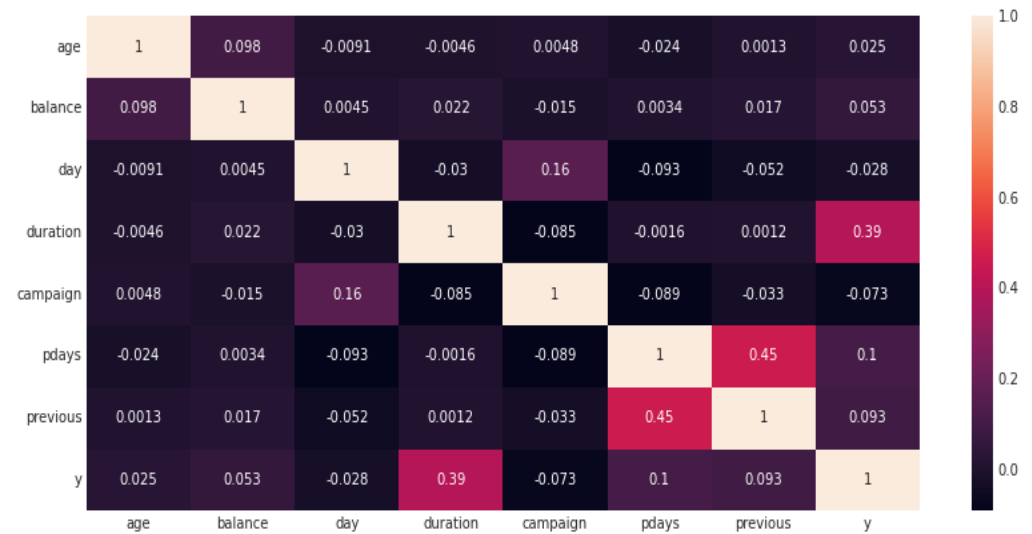
Data Understanding



Target variable ของเรา มีสัดส่วน 1 ต่อ 0 คิดเป็น 11.70 % ซึ่งเป็นตัวเลขที่ถือว่า สมเหตุสมผล และคิดว่าไม่ทำให้เกิด imbalanced data

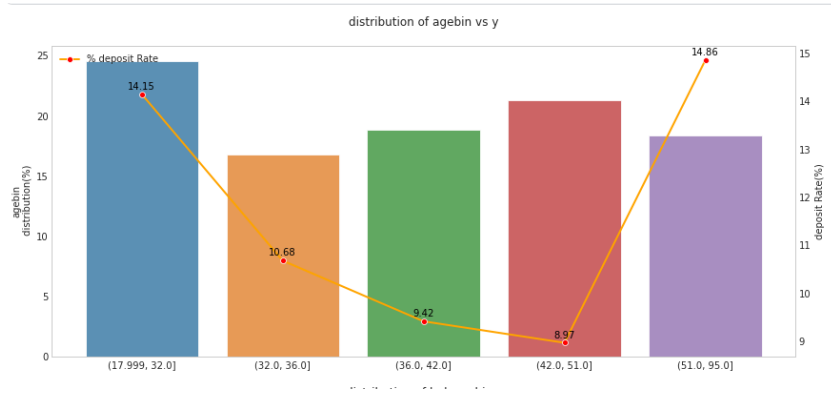


จากรูปจะเห็นได้ว่า ตัวแปรที่เป็น numeric โดยส่วนใหญ่นั้นจะเบ้ขวา อย่างมาก ทำให้สามารถตั้งสมมุติฐานได้ว่า พฤติกรรมโดยส่วนใหญ่ ของ data นี้ มีแนวโน้มไม่พทางจำนวนที่น้อย

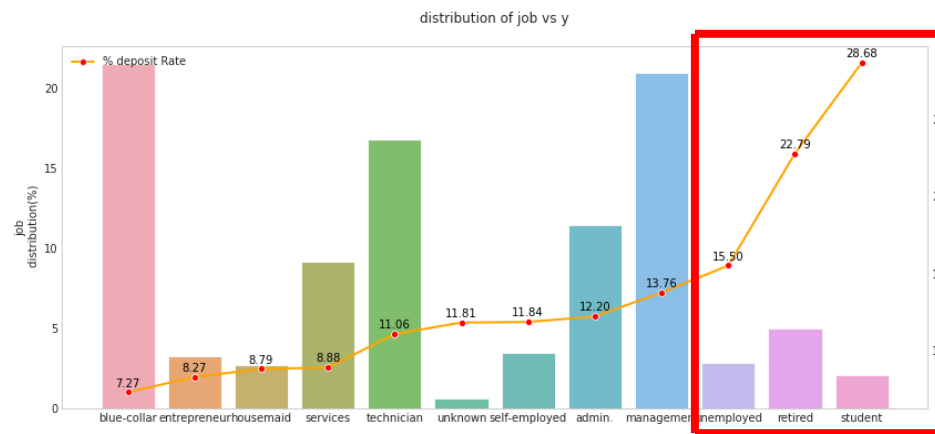


จากรูปจะเห็นได้ว่า y นั้นค่อนข้างมีความสัมพันธ์เชิงเส้นที่ดีกับ duration แต่ควรระวังเมื่อนำ duration เข้า ml model

EDA I : Age & Job

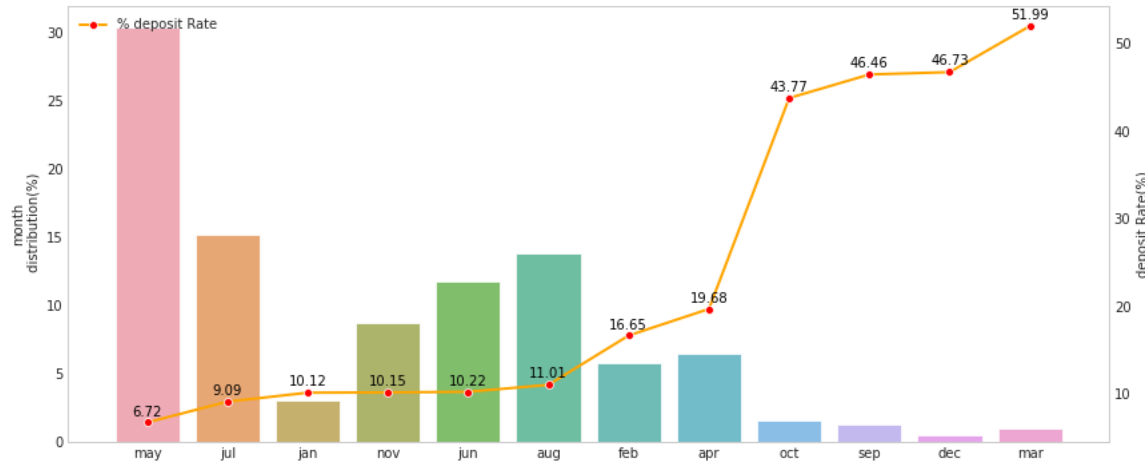


Insight : จากกราฟจะเห็นได้ว่า ช่วงอายุของคนที่จะฝากประจำ สูงอยู่ 2 ช่วงคือ ต่ำกว่า 30 ปีและ สูงกว่า 50 ปี ซึ่ง สอดคล้องกับอาชีพ ที่มีการฝากประจำเยอะที่สุด นั่นคือ นักเรียนและกลุ่มคนที่ รีไทร์



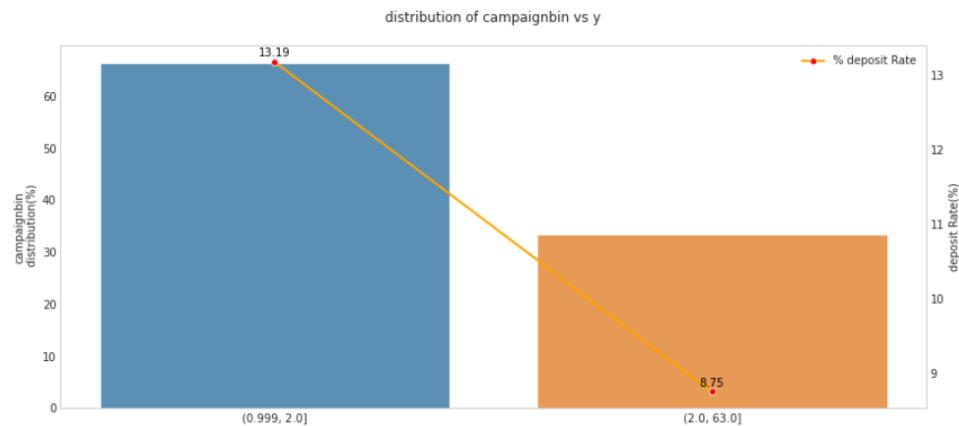
KEY : ดังนั้น หากเราจะเพิ่มทั้ง conversion rate และ sale volume เราต้องโฟกัสไปที่ กลุ่มเป้าหมายของเรา นั่นคือ วัยรุ่น และคนแก่

EDA II : Month & Campaign



Insight: จากกราฟจะเห็นได้ว่า อัตราการฝากประจำนั้น จะพุ่งสูงมากใน October ,September ,December และ march จากข้อมูลแสดงให้เห็นว่า ที่ผ่านมามาเราได้โฟกัสในการทำ marketing ในเดือน may มากเกินไป และผลลัพธ์ไม่ได้ออกมาดีเท่าที่ควร

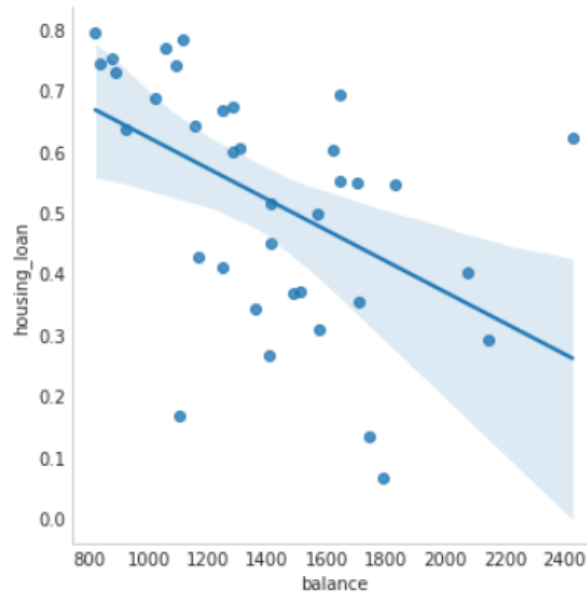
Key: ดังนั้น marketing campaign ต่อไปนั้น ควรจะจัดในเดือนที่มีการตอบรับสูง คือ October ,September ,December และ March



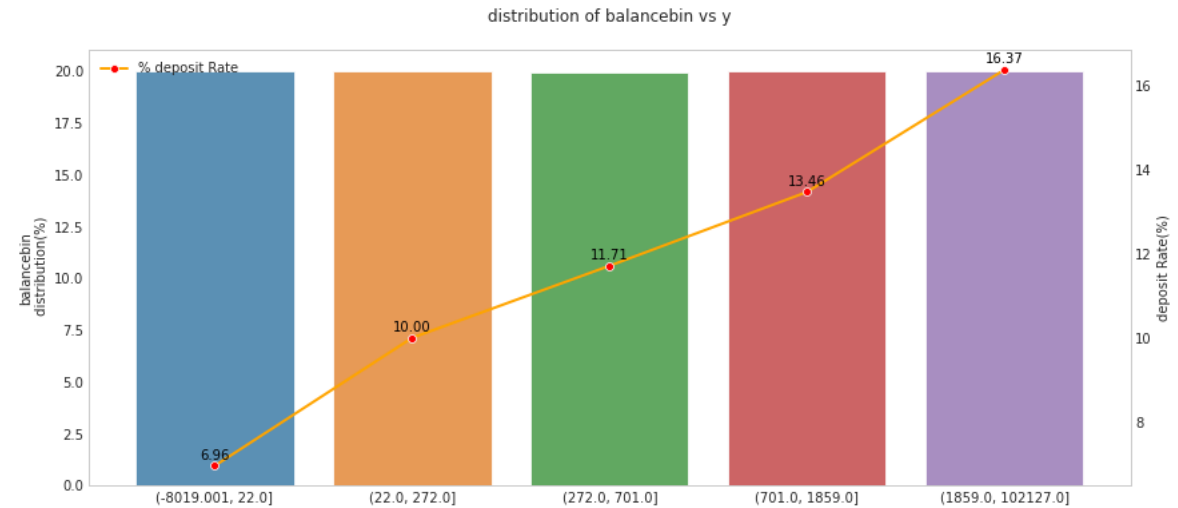
Insight: จำนวนการติดต่อลูกค้า นั้น มีผลอย่างมากต่อ user experience ลูกค้าไม่ชอบการติดต่อหลายครั้ง

Key: ดังนั้นหากพยายามโฆษณา หรือ ติดต่อ หรือ ขายตรงมากเกินไป จะทำให้ ลูกค้าไม่สนใจ campaign ของเรา และเสีย user experience ที่ดีไป ทางแก้อาจใช้เป็น Native Advertising แทน

EDA III : Balance & housing & loan

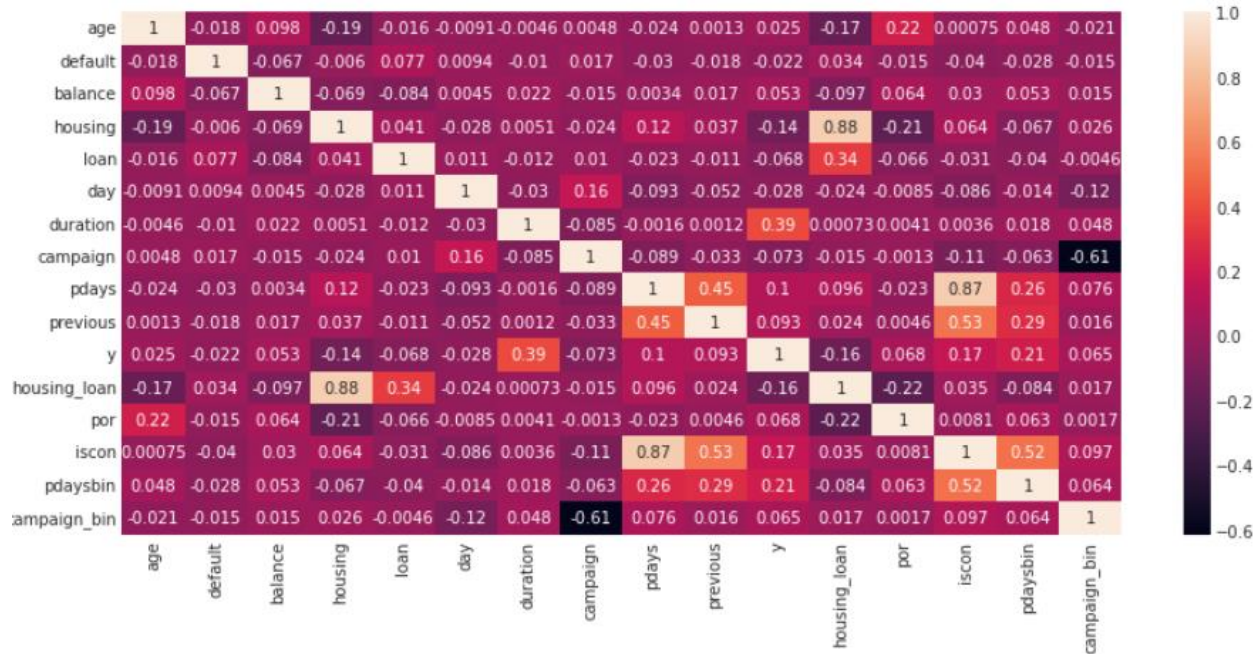


Insight : จากกราฟจะเห็นได้ว่า ผู้ที่มี balance ต่อเดือนสูงนั้น หมายความว่า จะมีแนวโน้มที่จะ ผากประจำ สูงขึ้น รายได้และ การกู้ยืม (ทั้ง housing loan และ personal loan มีความสัมพันธ์เชิงเส้น แบบผกผันกันอย่างชัดเจน การกู้ยืมนั้นเป็นปัจจัยสำคัญ เนื่องจาก การกู้ยืมโดยส่วนใหญ่นั้นต้องจ่ายเงินผ่อน หรือดอกเบี้ยในทุกๆเดือน ในกรณีที่ลูกค้ามีการกู้ยืม และมี balance ที่ต่ำ มีโอกาสน้อยมากที่ลูกค้าจะนำเงินมาฝากประจำ เพราะต้องใช้เงินทุกเดือน บวกกับ อาจไม่มีเงินเหลือพอให้ฝาก



Key : ดังนั้น campaign การตลาดนี้ต้อง คำนึงถึงผู้มีรายได้สูง หรือ มีการกู้ยืมน้อย ซึ่งสอดคล้องกับ potential customer ที่เป็นกลุ่ม และ คนชรา เพราะคนทั้ง 2 วัยนี้ ต้องการความเสี่ยงที่ต่ำ และ ไม่ได้มีการกู้ยืมมากนัก การได้ข้อมูลเหล่านี้มาทำให้เรา โฟกัสไปที่กลุ่มเป้าหมายได้มากขึ้น และ สามารถ personalize ประสิทธิภาพที่ potential customer แต่ละกลุ่มของเราจะได้รับ

Feature engineering



- pdaysbin สร้างจากการ EDA ซึ่งพบว่า ลูกค้าจะตอบรับ แคมเปญมากที่สุด เมื่อเสนอแคมเปญ ห่างจากวันล่าสุดที่ตอบรับล่าสุด 3 เดือน และ 6 เดือน
- por อัตราส่วนระหว่าง ค่าเฉลี่ยของ balance แต่ละ job/marital หารด้วย อัตราการกู้ยืม
- housing_loan มีการกู้ยืมหรือไม่ ทั้งบ้าน และ ส่วนตัว
- campaign_bin สร้างจากการ EDA ซึ่งพบว่า จำนวนการติดต่อ ต่อ 1 คนนั้น ไม่ควรเกิน 2 ครั้ง
- age_bin สร้างจาก การ eda ว่า potential customer ของเรานั้น คือคนอายุน้อยกว่า 30 ปี และคนอายุมากกว่า 60 ปี
- iscon เคยติดต่อลูกค้าคนนี้นั้มาหรือยัง

*Note : ข้อมูล ที่เป็น categorical data ได้ทำการแปลงโดยใช้ one hot encoding แล้ว แต่ไม่ได้นำมาแสดงในสไลด์ เนื่องจากมี feature เกิดขึ้น จากการทำ one hot เยอะมาก เกรงว่าจะไม่สามารถ อ่านออกได้โดยง่าย ใน heatmap จึงขอเสนอเพียง numeric data

Modeling and evaluation

Model XGBoostClassifier

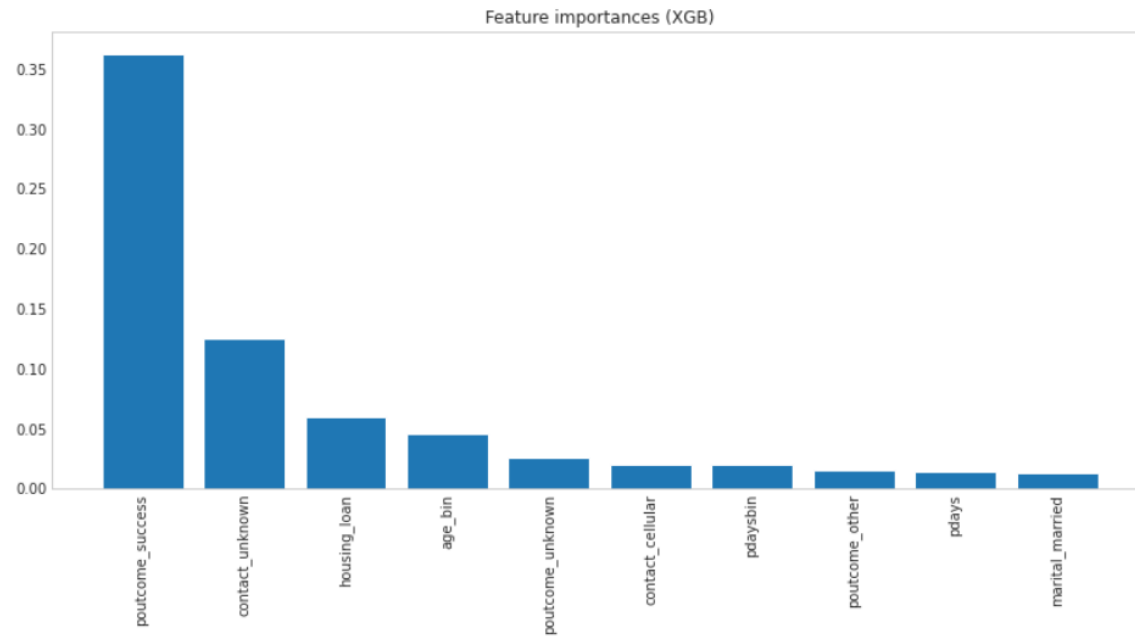
`xgboost.XGBClassifier(n_estimators=100, learning_rate=0.08, gamma=0, subsample=0.7, max_depth=7)`

Note. หากดูจาก heatmap ในหน้าที่ผ่านมา จะเห็นได้ว่า duration มี correlation กับ y สูงที่สุด แต่เราไม่สามารถนำมาใส่โมเดลได้เนื่องจาก duration เป็นข้อมูลที่จะมีการอัปเดตเรื่อยๆ หรือเรียกได้ว่าเป็นข้อมูลที่จะเปลี่ยนไปเมื่อเวลาผ่านไป หากนำมาใส่ จะเกิด data leakage ได้

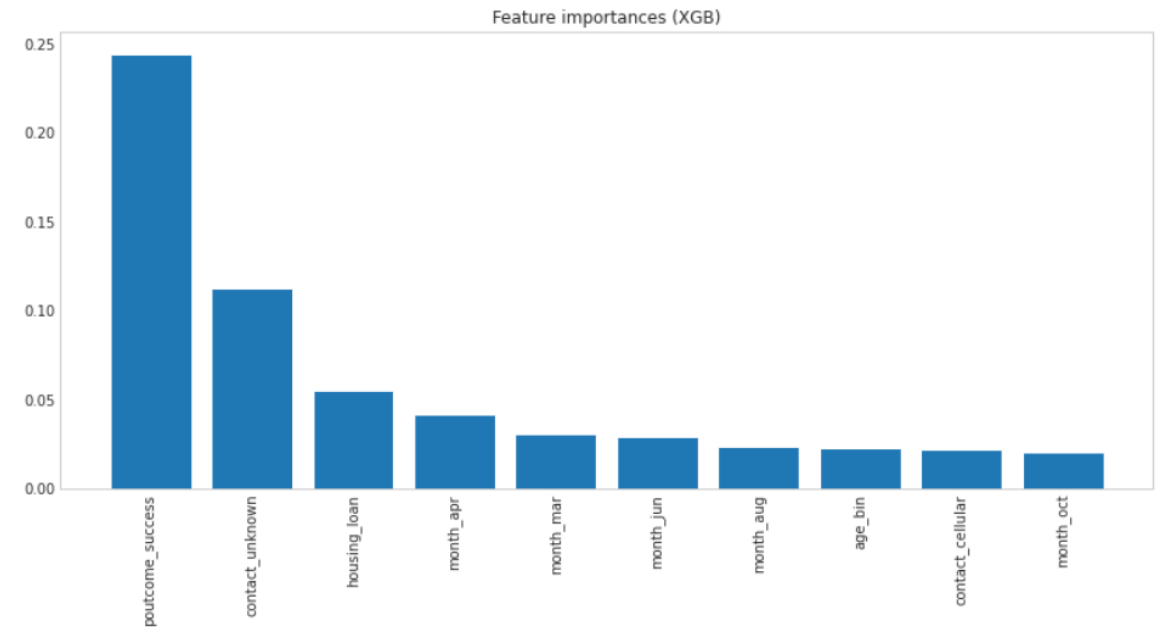
train		test	
roc_auc_score	0.896303374633579	roc_auc_score	0.7646106545165545
f1_score	0.41142115509409466	f1_score	0.2790905037895675
recall_score	0.881371640407785	recall_score	0.6285140562248996
precision_score	0.26834085778781036	precision_score	0.17936962750716332
accuracy_score	0.9101713380211944	accuracy_score	0.8916219839142091
On train set		On test set	

Conclusion : จากการวัดผล พบว่า โมเดล นั้นมี accuracy ที่น่าพอใจ แต่ f1 , recall ,precision นั้น ค่อนข้างต่ำ คิดว่า หากต้องการให้โมเดลมี precision มากกว่านี้ อาจทำได้โดยการลด threshold หรือเพิ่มข้อมูล

Top 10 Importance features



Do not include month feature



Include month feature

EDA

Feature Engineering

Modeling

Evaluation

Execution

From Analytics to Execution

จากการวิเคราะห์ข้อมูล หากเราต้องการเพิ่ม conversion rate และ sale volume เราจะต้อง แบ่งกลุ่มลูกค้าเป็น 3 cluster คือ

1. ผู้ที่มีอายุ น้อยกว่าเท่ากับ 30 ปี หรือ นักเรียน นักศึกษา
2. ผู้ที่มีอายุ มากกว่า 30 แต่ น้อย กว่า 55 ปี และมีรายได้สูง
3. คนชรา วัยเกษียณ (รายได้ปานกลาง ถึง รายได้สูง)

ทั้ง 3 กลุ่มนี้คือ potential customer ของเรา

กลยุทธ์เพื่อที่จะเพิ่ม conversion rate และ sales volume นั้น คือ เราจะต้องสร้างความต้องการขึ้น เพราะ term deposit นั้น อาจไม่ได้ร้อนแรงเหมือนแต่ก่อนแล้ว มา รายละเอียดดังนี้

1. ผู้ที่มีอายุ น้อยกว่าเท่ากับ 30 ปี หรือ นักเรียน นักศึกษา หากต้องการจะเปิดบัญชีฝากประจำ แสดงว่าคนกลุ่มนี้นั้น ยังไม่มีความต้องการใช้เงินมาก ต้องการออมเงิน เพื่อเก็บไว้ใช้ในอนาคต และ ยังไม่พร้อมที่จะรับความเสี่ยง
2. ผู้ที่มีอายุ มากกว่า 30 แต่ น้อย กว่า 55 ปี และมีรายได้สูง ถึงแม้จะเป็นช่วงวัยที่มีค่าใช้จ่ายเยอะที่สุด มีความพร้อมที่จะรับความเสี่ยง เงินฝากประจำอาจจะไม่เหมาะกับลูกค้ากลุ่มนี้ หากแต่เราสามารถเพิ่มความต้องการอยากจะทำฝากประจำได้โดย เล่นกับสถานการณ์ปัจจุบัน เพราะในโลกปัจจุบัน เศรษฐกิจผันผวน การลงทุนที่ให้ผลตอบแทนมากก็อาจขาดทุนมาก ถ้าเราให้ข้อเสนอ ในแง่มุมมองของ ความปลอดภัย ไร้ความเสี่ยง จะสามารถดึงดูดกลุ่มนี้ได้
3. คนชรา วัยเกษียณ (รายได้ปานกลาง ถึง รายได้สูง) คนในวัยนี้นั้น ค่อนข้างต้องการความเสี่ยงที่ต่ำ ไม่อยากลงทุนกับอะไรที่ผันผวน แต่เรื่องที่สำคัญที่สุดในวัยนี้คือ สุขภาพ หากเราให้ข้อเสนอในแนวของ หากฝากประจำกับ ธนาคาร จะได้ ประกันสุขภาพไปด้วย จะสามารถดึงดูดกลุ่มนี้ได้