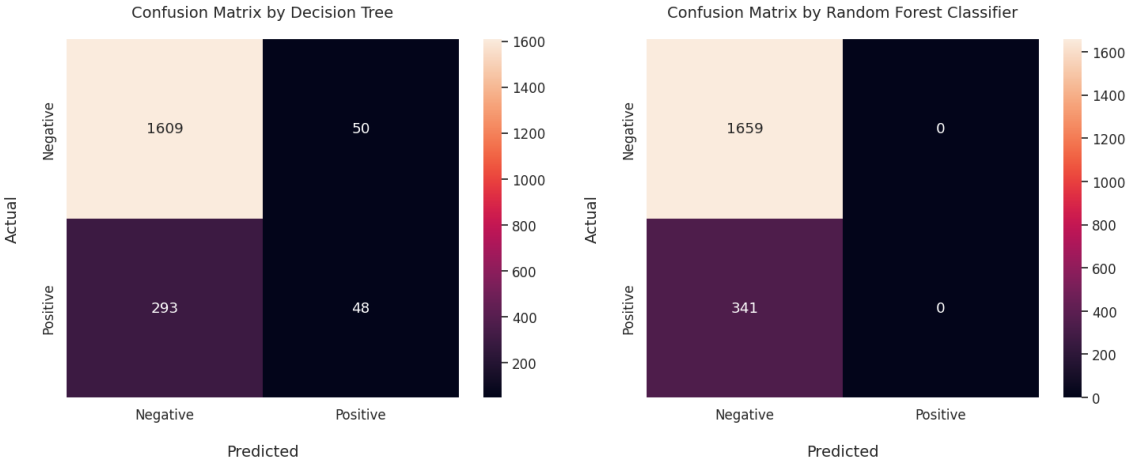
Week 9: (Homework) Decision Tree กิ่งรัก ไพเราะ 6230040421

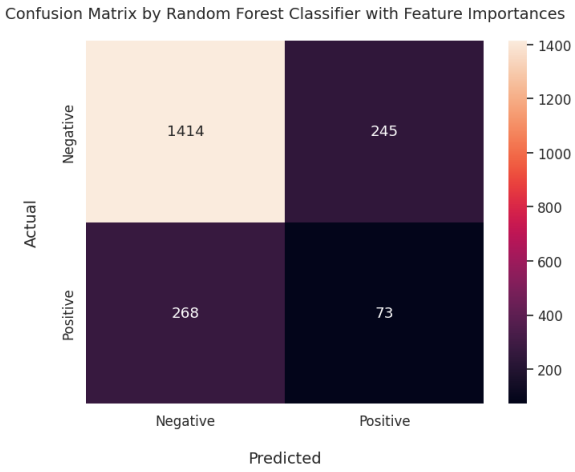
1 - 4b,4d Colab link:

https://colab.research.google.com/drive/1WJ1ZoeMWDL3YrMO2CFRz9Pf1x3tUPXnK?usp=sharing

4c. Find feature importance from Random Forest learning, then use the important feature subsets for training CART again. Compare the “metric” performance between the three models.



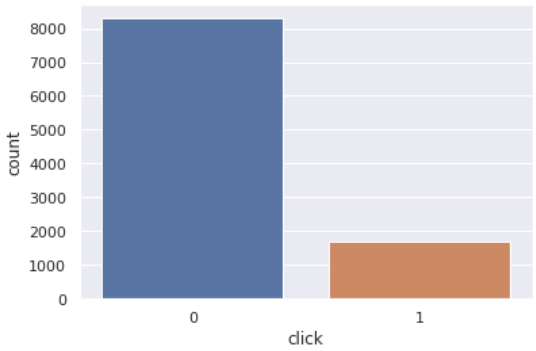

รูปที่ 1 และ 2 แสดง Confusion Matrix เมื่อใช้แบบจำลอง Decision Tree และ Random Forest โดยสุ่ม Hyperparameter ตามลำดับ




รูปที่ 3 แสดง Confusion Matrix เมื่อใช้แบบจำลอง Random Forest โดยสุ่ม Hyperparameter และใช้ Feature Importance

จากการเปรียบเทียบ Confusion Matrix ทั้งสามแบบจำลองจะเห็นได้ว่าจำนวนที่ทายผิดและถูกของการทายเป็น 0 (Negative) ทุกแบบจำลองค่อยข้างมีค่าที่ใกล้เคียงกัน ส่วนของการทำนายเป็น 1 (Positive) มีความต่างกันเล็กน้อย โดยเมื่อใช้ Feature Importance ในแบบจำลอง Random Forest พบว่าทำให้การทำนายเป็น 1 มีประสิทธิภาพมากขึ้น

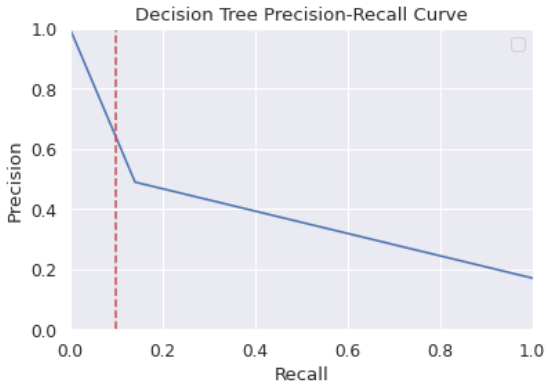
4e. Please explain why this “metric” makes more sense that the conventional accuracy in the click-through prediction task.

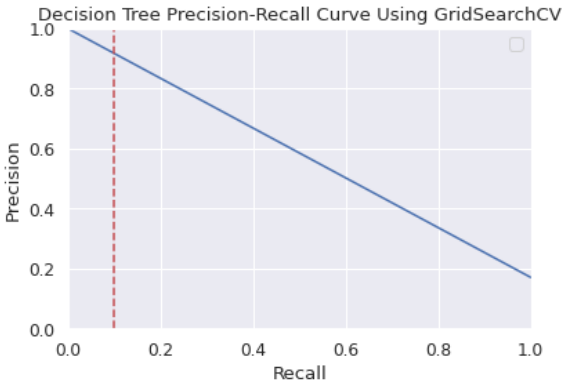



รูปที่ 4 แสดงจำนวนทั้งหมดของแต่ละ target class

จาก Confusion Matrix ในรูปที่ 1-3 จะเห็นได้ว่าทุกแบบจำลองจะมีจำนวนที่ทายผิดและถูกของการทำนายเป็น 0 (Negative) หรือ False Negative และ True Negative ที่มากกว่าการทำนายเป็น 1 (Positive) หรือ False Positive และ True Positive เนื่องจากเมื่อสังเกตปริมาณข้อมูลใน Data Set ในรูปที่ 4 พบว่าปริมาณ Target Class ของ 0 มากกว่า 1 มากจึงส่งผลให้ผลการทำนายผิดและถูกของการการทำนายเป็น 0 มีปริมาณมากกว่า

5. In CART, students are encouraged to tweak some hyperparameters, such as max_depth, min_samples_split, with GridSearchCV(). What is the highest “precision at recall 0.1” you are able to achieve?

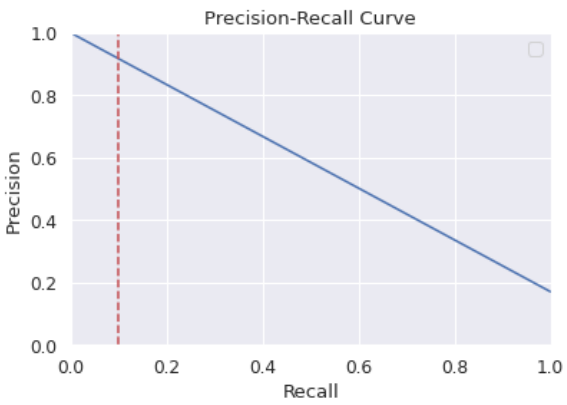


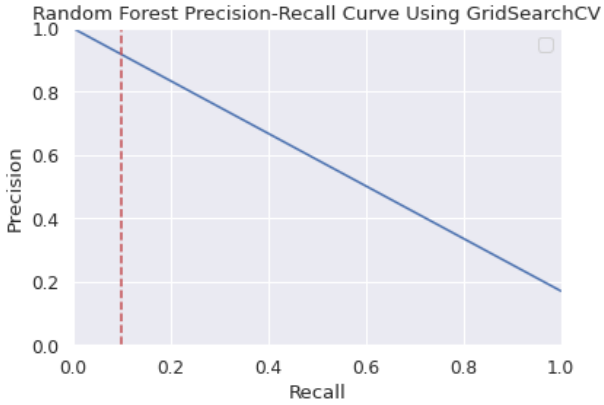




รูปที่ 5 และ 6 แสดง Precision-Recall Curve ของผลการทำนายแบบจำลอง Decision Tree โดยการสุ่ม Hyperparameter( max_depth = 10, min_samples_split = 10) และใช้ GridSearchCV ( max_depth = 2, min_samples_split = 2) ตามลำดับ

จากรูปที่ 5 จะได้ว่าหากกำหนด Recall = 0.1 จะได้ Precision ≈ 0.65 แต่เมื่อใช้ GridSearchCV ในรูปที่ 6 ที่ Recall = 0.1 จะได้ Precision เพิ่มขึ้นถึงประมาณ 0.9

6. In Random Forest, you can also tweak hyperparameters, such as min_samples_split, max_features, and n_estimators. What is the highest “precision at recall 0.1” you are able to achieve?




รูปที่ 7 และ 8 แสดง Precision-Recall Curve ของผลการทำนายแบบจำลอง Random Forest โดยการสุ่ม Hyperparameter ( max_depth = 5, min_samples_split = 5, n_estimators = 10 ) และใช้ GridSearchCV ( max_depth = 2, min_samples_split = 7, n_estimators = 1 ) ตามลำดับ

จากรูปที่ 7 และ 8 ที่ Recall = 0.1 จะได้ Precision ประมาณ 0.9