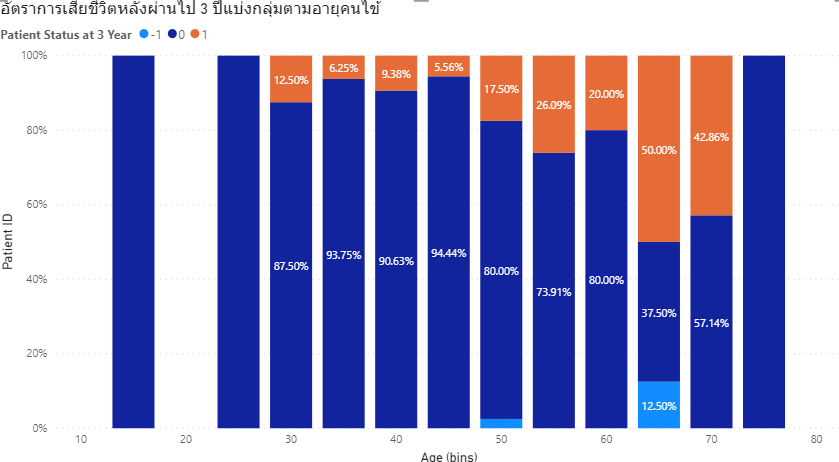
แน่นอนว่าจากกลุ่มตัวอย่าง (ที่ตัด unknown ออก) มีความห่างกันของ label ทั้งสองกลุ่มมากคือมีผู้รอดชิวิตหลัง 3 ปี 164 คนและ เสียชีวิตก่อน 3 ปี 31 คน อัตราส่วนของผู้รอดชีวิตหลัง 3 ปี คือ 85% ทำให้ accuracy น่าจะไม่ใช่ตัวชี้วัดที่ดีนัก ในกรณีนี้ เพราะการเดาสุ่ม โดยปกติก็มีความแม่นยำ 85% แล้ว

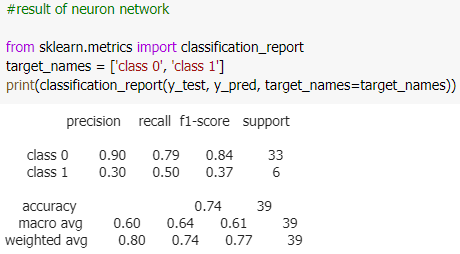
ในฐานะคนไม่มี Domain Knowledge แล้วคิดว่าในกรณีนี้ เรามุ่งเน้นกับ recall rate เพราะอะไร? เพราะนี่คือการ ทำนายผลเพื่อใช้ในการตัดสินใจในการรักษา เราต้องการให้ model จับคนที่มีแนวโน้มจะเสียชีวิตภายใน 3 ปีให้มากที่สุด เพื่อทำการดูแลรักษาอย่างใกล้ชิด นั่นจึงทำให้ Type II error หรือกรณีที่คนไข้มีแนวโน้มจะเสียชีวิตแต่ model ทำนายผลว่าไม่ ก่อความเสียหายมากกว่า Type I error หรือกรณีที่คนไข้ไม่น่าจะเสียชีวิต แต่ model ทำนายผลว่าเสียชีวิต เพราะอย่างมากก็แค่ดูแลรักษาใกล้ชิดขึ้น ซึ่งไม่เสียหายอะไร (ทั้งนี้ทั้งนั้น ถ้ามี Domain Knowledge ทางการแพทย์เข้ามาช่วยก็จะทำให้การตัดสินใจทาง model สามารถทำได้ดีขึ้น แต่เราต้องการนำเสนอ model ว่าในฐานะคนเลือก model เราเลือกที่จะ minimize Type I error และ maximum recall rate เพื่อให้คนไข้ที่มีความเสียงจะเสียชีวิตสูงกว่าคนไข้คนอื่น ได้รับการดูแลรักษาอย่างใกล้ชิด)

ถัดมาคือปัญหาเรื่องการ เลือก Algorithm เราคิดว่าในกรณีนี้เราเลือกใช้ neuron network เพราะเราไม่ต้องการความสามารถในการอธิบายของ model ตัว feature โดยส่วนใหญ่ extract มาจากภาพฟิลม์ของคนไข้ ซึ่ง neuron network น่าจะเป็น algorithm ที่คนเลือกใช้ และตอบโจทย์ปัญหาในเชิง image classification มากที่สุด (ทั้งนี้ถึงแม้ว่าในทางความเป็นจริงแล้วนี่ไม่ใช่ปัญหาเรื่อง image classification ซะทีเดียวเพราะจากการ explore data เราพบว่า pattern การเสียชีวิตหลัง 3 ปีมันมีแนวโน้มสูงขึ้นตามอายุของคนไข้เสียมากกว่า)

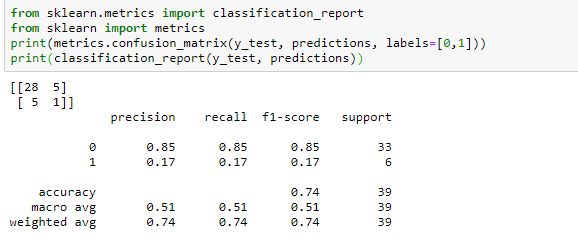
(จากภาพจะเห็นว่าอัตราการเสียชีวิตของคนไข้มี pattern ที่มากขึ้นค่อนข้างชัดเจน ตามอายุมากขึ้น)



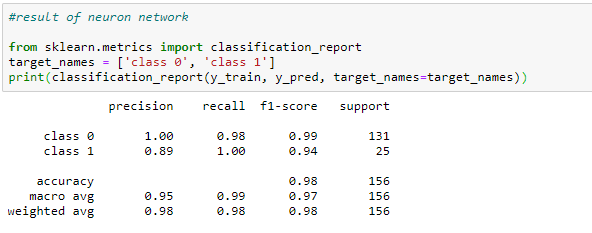
แต่เราคิดว่าในมุมมองของเราโจทย์ให้ feature เกี่ยวกับ จุดของ image มาจึงต้อง assume ว่าต้องใช้ feature ของ image ให้เกิดประโยชน์ เราเลยรัน neuron network ดู โดยตัด feature ที่ไม่เกี่ยวกับ ลักษณะของคนไข้ออก (PatientID,Time to event,Patient Status) และ remove คนไข้ที่เราไม่รู้ status หลัง3ปี ออกไปด้วย (ดูโมเดลได้ในโค้ด jupyter notebook)

ซึ่งจากภาพข้างล่างแสดงผลการ run neuron network ด้วย threshold ที่ 0.05 ซึ่งจะทำให้ได้ recall rate ที่สูงที่สุดโดยไม่สูญเสีย precision มากเกินไป (คือมีค่า f1-score ที่ดีทั้งสอง class) แต่เรามุ่งเน้น recall ของ class 1 นั่นแสดงให้เห็นว่า model สามารถจับ คนไข้ที่มีแนวโน้มจะเสียชีวิตได้ถึง 50% ซึ่งเป็นตัวเลขที่ไม่แย่ ถ้าเทียบกับว่าในคนไข้ทั้งหมดมีแค่ 15% เท่านั้นที่เสียชีวิตหลังผ่านไป 3 ปี 

ทั้งนี้ทั้งนั้น เราได้ลองรัน random forest ดูด้วย เพื่อใช้เป็น baseline ในการเลือก model ซึงจากภาพด้านล่าง จะเห็นว่าแม้ accuracy จะไม่ต่างกันแต่ recall rate class 1 ของ neuron ด้านบนดีกว่ามาก (0.50 ต่อ 0.17) ซึ่งเป็นจุดประสงค์ของผู้ทำ model ที่ต้องการมุ่งเน้นจุดนี้ เราจึงไม่เลือกใช้ algorithm random forest และใช้ neuron ที่ threshold 0.05



ปัญหาของ model

อย่างไรก็ดีปัญหาหลักของ model ก็คือมัน overfit มาก จากภาพด้านล่างมันแทบจะทำนาย train set ได้แม่นยำสุดๆ (ซึ่งเกิดขึ้นได้ในกรณีที่ data เป็น linear แต่มันกลับ drop ลงเมื่อมันต้องทำนาย data ที่มันไม่รู้จักอย่าง test set (หรือ validate set) คิดว่าเป็นเพราะ data มันน้อย และเราก็ไม่อยาก down sampling imbalance data ด้วยเพราะ data มันน้อยอยู่แล้ว 

* Explore variability in radiomic features (v1-v482) using the data from 30 patients.
* Address the issue of variability feature values in your model performance evaluation.

จากข้อมูลจะเห็นว่าคนไข้แต่ละคน มีค่า feature ที่แตกต่างกันอันเนื่องมากจากแพทย์ 3 คน Annotate ROI ของแต่ละ Feature ต่างกัน (ซึ่งใน dataset ทั้งหมด เอา ค่า ROI มาจากแพทย์คนที่ 3 ) ซึ่งทำให้ model performance อาจมีผลที่แตกต่างกันไปขึ้นอยู่กับแพทย์ที่ส่ง input

ปัญหาเรื่องนี้แก้ง่ายมาก ซึ่งเป็นเคสเดียวกับกรณีที่ facebook ใช้ classification comment ทางทีดีและแย่ ซึ่งในเมื่อค่า ROI นี้แตกต่างกันตามแพทย์ที่ draw ค่า ผมจึงเสนอให้ใช้ median ของแพทย์ทั้ง 3 คนเป็น feature โดยอาจให้แพทย์ 3 คนขึ้นไปเป็นคน ส่งค่า ROI ให้ model และใช้ค่า median ของแต่ละ feature เป็นตัว input ของ model

ทำไมถึงเป็น median ? เพราะว่า median เป็นค่ากลางทางสถิติที่ไม่แปรผันตาม extream value,outlier, และ human error กล่าวก็คือสมมติมีแพทย์ซัก 1 คน draw ค่า ROI ผิด เพิ่ม 0 ขึ้นมาข้างหลัง 1 ตัวทำให้เกิด outlier ค่า median ก็ยังไม่เปลี่ยนแปลงซึ่งต่างกับค่า เฉลี่ยหรือการเลือกเอาค่าจากแพทย์คนใดคนหนึ่งเพราะอาจเกิด human error หรือ bias ขึ้นได้ ซึ่งในกรณีที่ค่า input feature ของ model มันเกิดจาการ estimate ซึ่งแปรผันไปตามแต่ละบุคคล การใช้ค่า median ทำให้ model มีความเสถียรมากขึ้นลดความเสี่ยงจากข้อมูล noise ,bias หรือข้อมูลที่เกิดจาก human error ไปได้