Towards Better Insights: Preliminary Results Of A Machine Learning Model For Fracture Risk Assessment

Features used in Osteoporosis Machine Learning Models

Alex Canales, Mohit Doel, Volodymyr Voytenko
2021/09/27

Selecting features (Columns) for machine learning is done in 3 steps:

1. An assortment of features are selected using information obtained from academic research and specialist recommendations. *(This is the largest set of features)*

2. Features are then removed or swapped depending on how much they influence the final machine learning predictions. This is called **feature importance**. For example, age had a major influence on the final prediction, but smoke did not. Therefore, we remove the smoke column, but keep the age column. *(This is the second largest set of features) We repeat this part 2 times!*

3. The process of **feature importance** keeps repeating until there is no more substantial gain. The remaining features are thrown into a list and called the **final feature set**. *(This is the smallest set of features)*

Please take note! The machine learning models select features in different ways. For example, the random forest model removes the smoke column (Smoke was considered not useful), in contrast, the linear regression model kept the smoke column.

Osteoporosis research such as FRAX and CAROC have already shown which features are the most important, but this does not mean the models rank the features the same way For example, Osteoporosis research has shown that gender is more important than weight because women have a much higher chance of developing osteoporosis than men; but some machine learning models, such as decision trees, weight was seen as more important than gender. This is because the weight value had a greater influence on the final prediction.

Because of these differences in models, we are only going to focus on our **best performing** algorithm. Polynomial with Degree 2.

All Features (Full Set)

This is a list of all the features that were used in the Machine Learning model. There are 29 features in total:

1. Gender
2. Age
3. Weight
4. Height
5. Parent Break
6. Arthritis
7. Cancer
8. Patient Unsteady
9. Where Live
10. Education
11. Diabetes
12. Heart Disease
13. Respiratory disease
14. Alcohol
15. How break (How the bones were broken)
16. Hip
17. ankle
18. clavicle
19. elbow
20. femur
21. spine
22. wrist
23. shoulder
24. tibfib
25. Was fract due to fall
26. Patient Fall
27. Patient worried (fxworried)
28. Patient not working (notworking)
29. Patient marital status (marital)

Many of these features did not influence the final prediction. Although some features such as Patient Fall did influence the final prediction, the effects were rather small and insignificant (Less than 1% Importance)

Selected Features 1 (Middle Set)

We narrowed down the feature set using feature importance. In total, 12 features remained.

1. Gender
2. Age
3. Weight
4. Height
5. Parent Break
6. Arthritis
7. Cancer
8. Patient Unsteady
9. Where Live
10. Education
11. Heart Disease
12. High Blood Pressure

Take Note! FRAX makes use of the #1,2,3,4,5,6,7 features. Smoke, Alcohol, and secondary osteoporosis did not make it to this set.

Maybe the Smoke, and Alcohol columns in FRAX should be replaced with Patient Unsteady, Where live, Education, Heart Disease, and the High Blood Pressure columns. Interestingly enough, Heart Disease and HBP was seen as more influential than Smoke and Alcohol in all the machine learning models we tested.

Selected Features 2 (Semi-Final Set)

We narrowed down the feature set using feature importance. In total, 7 features remained.

1. Gender
2. Age
3. Weight
4. Height
5. Parent Break
6. Arthritis
7. Cancer

Take Note! All of these Features are found in FRAX. Smoke, Alcohol, and secondary osteoporosis did not make it to this set. The Patient Unsteady, Where live, Education, Heart Disease, and the High Blood Pressure columns did not make it into this set either.

Best Features (Final Set)

The best features were added into the final set. In total, 4 features remained.
1. Gender
2. Age
3. Weight
4. Height

These 4 features had the greatest influence on the final prediction.
Take Note! CAROC also makes use of Gender and Age, but not Weight and Height. Maybe CAROC should include these features in the future.